U.PORTO

UNIVERSIDADE DO PORTO

2.º CICLO

FCUP
2016

U.PORTO

Estudo e Modelação de Informação Geográfica

Vanessa Cristina Azevedo Boucinha

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTC

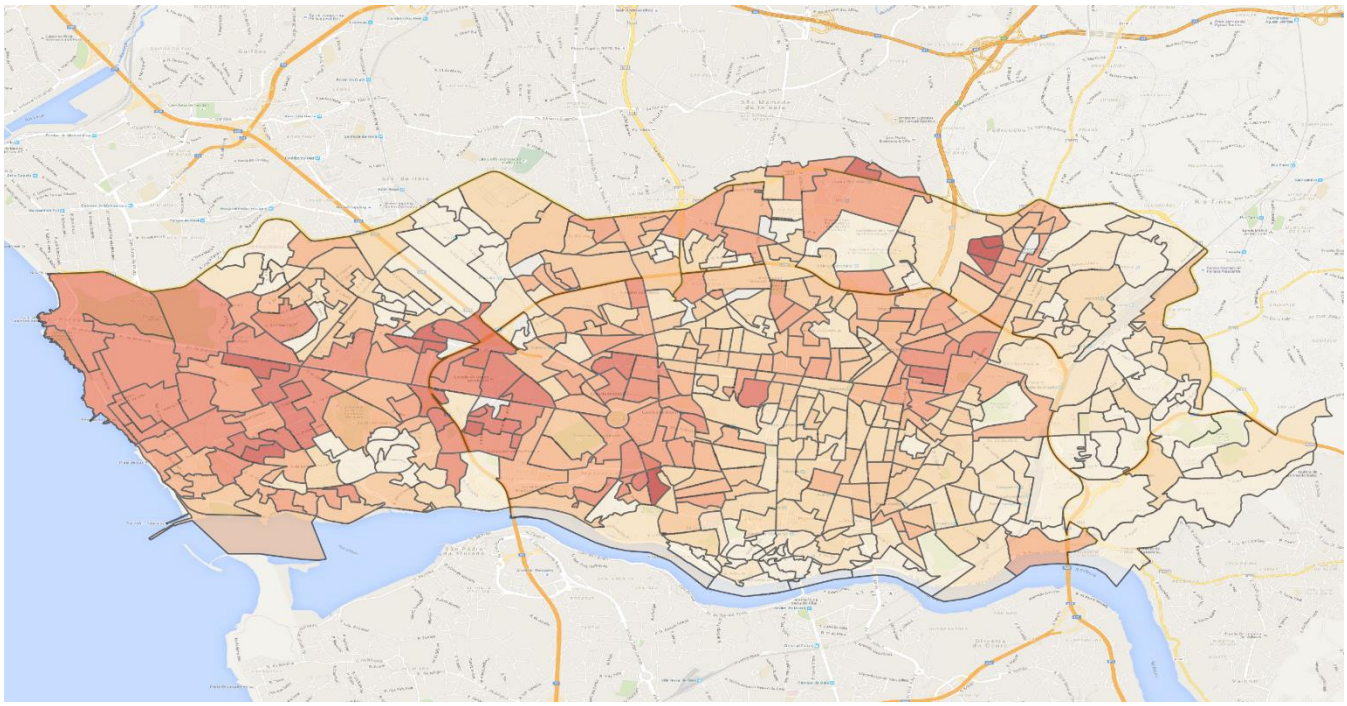
MSC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTC

Estudo e Modelação de Informação Geográfica

MSC

Vanessa Cristina Azevedo Boucinha
Dissertação Apresentada
à Faculdade de Ciências da Universidade do Porto em
Engenharia Física



Estudo e Modelação de Informação Geográfica

Vanessa Cristina Azevedo Boucinha
Mestrado Integrado em Engenharia Física
Departamento de Física e Astrofísica
20016

Orientador
Dr. João Manuel Viana Parente Lopes, Professor Auxiliar Convidado
Faculdade de Ciências

Coorientador
Dr. Filipe Pacetti, Consultor Sénior
Dr. Valeriy Brazhnyy, Consultor Sénior
Deloitte

Deloitte.

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Com estas palavras pretendo agradecer a todos aqueles que me ajudaram definir a minha personalidade bem como o meu percurso profissional e em especial este estágio.

Em primeiro lugar gostava de agradecer ao meu orientador, Dr. João Manuel Viana Parente Lopes, pela supervisão, pelas muitas horas gastas a auxiliar-me, pelo apoio em todos os momentos mais difíceis, por me dar sempre uma palavra de incentivo e ainda por tudo que me ensinou durante este período.

Quero ainda agradecer aos dois orientadores da Deloitte, Dr. Filipe Paccetti Correia e o Dr. Valeriy Brazhnyy, por me proporcionarem este desafio, que foi sem dúvida uma ótima experiência profissional, bem como pela disponibilidade e por tudo que me ensinaram no decorrer desta etapa.

Quero agradecer ainda à Faculdade de Ciências, todos os professores e amigos que passaram pelo meu caminho e me marcaram de alguma forma. Durante este percurso aprendi muito do que sei hoje em dia e acredito que o conhecimento que me passaram irá ser indispensável para conseguir ter uma carreira de sucesso.

Por fim quero agradecer a toda a minha família e ao meu namorado, que me ajudaram durante todos estes anos a superar os momentos mais difíceis e a nunca desistir. Sem eles este percurso não era possível.

Este projeto, em parceria com a empresa Deloitte consultores SA, tem como objetivo o desenvolvimento de um sistema capaz de efetuar o estudo e modelação de dados geográficos de interesse para o sector financeiro.

No decorrer deste trabalho utilizam-se três fontes de dados. Uma delas é inicialmente disponibilizadas e para as restantes é necessária a criação de um processo de recolha. Grande parte do trabalho efetuado consiste no tratamento dos dados, combinando diferentes fontes de informação e adequando a melhor forma de representação possível às necessidades.

Para desenvolver este trabalho é fundamental explorar as bases necessárias para o desenvolvimento de um Sistema de Informação Geográfica (SIG), adquirindo conceitos geográficos que se prendem com os diferentes tipos de projeções e com as camadas distintas com as quais podemos trabalhar. São exploradas ferramentas tais como Python e QGIS (quantum geographic information system).

Adquiridos os conceitos necessários é possível desenvolver um SIG com aplicação no sector financeiro partindo de informações sobre uma dada entidade bancária e outras recolhidas através de fontes livres.

This project, in partnership with Deloitte consultors SA, is aimed at developing a system that is able to study and model geographic data of interest to the financial sector.

During this work three data sources were used; one of them is initially made available and for the others a collection process to be provided. Much of the work done is based on the processing of data by combining different sources of information and adapting the best possible representation to the needs.

In order to carry out this work, it is fundamental to explore the basis needed for the development of a Geographic Information System (SIG), acquiring geographical concepts which relate to the different types of projections and the different systems that can be worked with. Tools such as Python and QGIS (quantum geographic information system) are explored.

After acquiring the necessary concepts, a SIG application can be developed in the financial sector on the basis of information about a given banking institution and other collected from open sources.

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

SIG - Sistema de Informação Geográfica
- GIS - Geographic Information System
- QGIS - Quantum Geographic Information System
- CTT - Correios, Telégrafos e Telefones
- OSM -Open Street Maps
- GAPI - Aplicação Google Maps
- BING - Bing Maps
- WGS84 - World Geodetic System 84
- ETRS89 - European Terrestrial Reference System 89
- INE - Instituto Nacional de Estatística

1 Introdução

Este relatório refere-se ao estágio realizado na empresa Deloitte consultores SA no departamento de analytics, em sistemas de informação geográfica (SIG). No âmbito do estágio, pretendeu-se explorar as potencialidades dos SIG no apoio à gestão da rede de agências de uma instituição financeira em Portugal.

O desenvolvimento da tecnologia é um grande aliado das empresas de consultoria. Nesta indústria o produto transacionado é o conhecimento. Por este motivo é decisiva a constante atualização para se disponibilizar as melhores soluções do mercado. Só fornecendo a solução mais otimizada aos seus clientes, a Deloitte vai ao encontro das suas expectativas. A introdução da tecnologia de SIG na empresa é bastante recente e este estágio pretendeu fortalecer esta especialização e desenvolver o seu potencial.

Partindo de um ficheiro com informações sobre cada cliente, de entre as quais consta a morada devidamente codificada, pretendeu-se que neste período fosse realizado um SIG capaz de efetuar uma análise geográfica de diferentes dados. O SIG pode dar uma informação valiosa para a gestão da rede de balcões de uma entidade financeira uma vez que permite a visualização toda a informação em torno de um balcão ao nível geográfico. Para a realização deste projeto é necessário perceber como uma entidade financeira funciona e a importância que um sistema SIG pode ter na sua otimização.

Neste capítulo introdutório iremos explorar a estrutura e a atividade das empresas de consultoria, mantendo o foco na Deloitte. De seguida, enquadraremos a atividade no sector financeiro. Por fim iremos introduzir o conceito de Sistema de Informação Geográfica e dos seus componentes, sendo este o principal tema deste projeto.

1.1 Empresas de consultoria

Diferentes entidades recorrem a serviços de uma empresa de consultoria quando a criação ou formação de recursos humanos internos, para tarefas tão especializadas, não se justifica. Os serviços prestados têm como propósito aumentar a eficácia do negócio e a produtividade da empresa.

Contratar uma empresa de consultoria tem vantagens e algumas desvantagens. Podemos identificar como desvantagem o facto do funcionário interno (da entidade que contrata o serviço) possuir um maior conhecimento da empresa, dos seus procedimentos e da complexidade dos problemas a resolver. Por outro lado, como o funcionário está limitado a uma só realidade, mesmo tendo conhecimento teórico sobre o problema, em geral, falta-lhe a experiência da implementação da solução. Contratar um consultor externo implica quase sempre que este tem um menor conhecimento da empresa. No entanto, em geral, o consultor externo tem uma experiência muito maior, pois já resolveu casos semelhantes noutras indústrias/empresas e

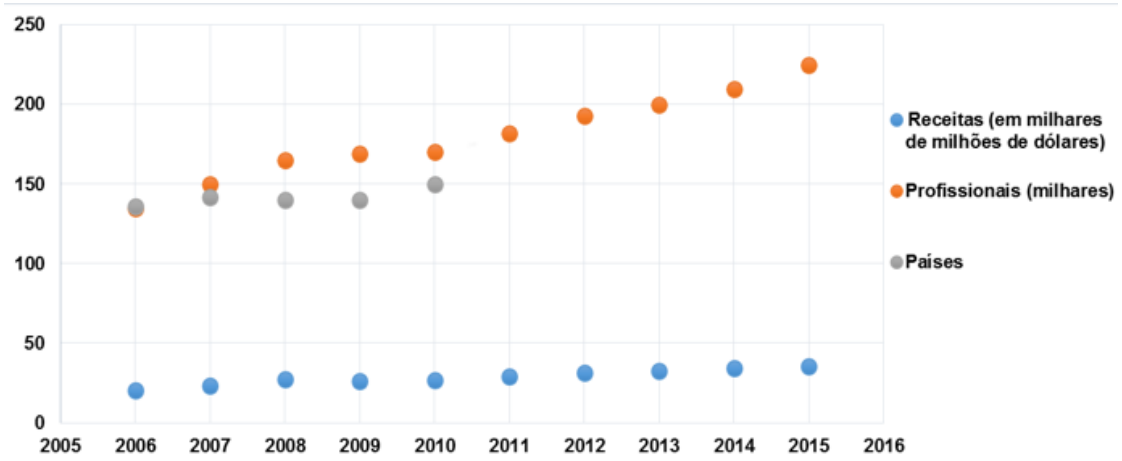


Figura 1.1: Dados de receita, número de profissionais e países da Deloitte.

teve a oportunidade de perceber os resultados obtidos. Outro fator que deve ser levado em conta, é que existem constrangimentos humanos na reestruturação das empresas aos quais um consultor externo é muito menos suscetível do que uma equipa interna.

Existe uma gama muito alargada de serviços prestados; podemos referir a área técnica relacionada com as características operacionais da atividade e a área de suporte à tomada de decisão da gestão. O trabalho a desenvolver num determinado projeto, é fortemente condicionado pelo problema inicial.

Na carreira de consultor pretende-se analisar e compreender transversalmente a informação disponível com o objetivo de fornecer as ferramentas necessárias ao exercício da gestão. Os serviços de consultoria têm como objetivo o diagnóstico dos problemas presentes nas empresas. É necessário perceber as necessidades do cliente e identificar soluções para problemas presentes assim como recomendações futuras. Neste tipo de atividade é comum existirem projetos a partir de problemas identificados em trabalhos anteriores. Um consultor deve tentar perceber todas as necessidades do cliente, mesmo que não se enquadrem no projeto a desenvolver no momento, podendo assim criar futuras propostas/soluções a implementar posteriormente.

1.1.1 A Deloitte

A Deloitte Touche Tohmatsu Limited, também conhecida como Deloitte é uma marca constituída por várias empresas, que se encontram em mais de 150 países e trabalham de forma independente. Trata-se de uma sociedade privada de responsabilidade limitada do Reino Unido e conta com mais de 225000 profissionais. É possível verificar através do gráfico da figura 1.1 que o número de profissionais, a receita e o número de países em que a Deloitte está representada estão a aumentar. Quanto ao número de países só se obteve acesso aos dados até 2010, sendo que nos anos seguintes este número manteve-se acima dos 150 mas não se encontra definido o número exato. [18]

Cada uma das empresas opera de forma independente e é responsável pelos seus atos e omissões. Quanto à Deloitte Portugal também tem seguido a tendência de crescimento sendo que, desde 2005, o único ano em que não se verificou crescimento da receita foi o ano de 2009, devido à crise ocorrida nesse período.

Esta empresa presta serviços em áreas distintas, sendo elas:

- Auditoria: O processo de auditoria tem o intuito de verificar a veracidade das demonstrações financeiras e não financeiras de uma empresa, proporcionando uma análise cuidada e sistemática das atividades da mesma. Este processo é efetuado através da análise de livros, contas e registos legais da organização. As auditorias têm-se vindo a evidenciar na prestação de serviços de elevada complexidade e importância no setor financeiro, demonstrando eficácia no auxílio às tomadas de decisão. A Deloitte apresenta um abrangente conjunto de profissionais de auditoria e serviços de risco capazes de colaborar com os clientes em serviços de consultoria e aconselhamento focando-se nos objetivos de negócio e melhoria de processos da empresa.

- **Consultoria**: A Deloitte distingue-se pela capacidade que apresenta no apoio eficiente à resolução dos mais complexos problemas apresentados pelos seus clientes. Os profissionais que contribuem para esta área evidenciam-se por serem capazes de desenvolver projetos desde a conceptualização das ideias até à implementação. Para isso disponibilizam um vasto conjunto de colaboradores combinando os mais distintos talentos e áreas de competências das quais se evidenciam Recursos Humanos, Estratégia e Operações e Tecnologia.

- Consultoria financeira: Os profissionais Deloitte em consultoria financeira prestam serviços de apoio e aconselhamento financeiro para clientes que pretendem evoluir o seu negócio quer por via de aquisições ou de expansões orgânicas. É desta forma fornecido um serviço que combina profundos conhecimentos das indústrias com a experiência quer em mercados locais assim como internacionais.

- Gestão de risco: As empresas parceiras Deloitte apresentam aptidão no sentido de implementarem estratégias e estabelecerem estruturas que potenciam a gestão de risco necessário. Por vezes o risco não tem de ser eliminado na sua totalidade, é necessário no entanto ponderar o risco que queremos admitir. Neste processo é preponderante o apoio fornecido pelos profissionais Deloitte para o desenvolvimento de abordagens estratégicas que oferecem soluções na gestão de risco e potenciam a criação de valor acrescentado respeitando sempre o cumprimento de regulações.

Os serviços de consultoria podem ser divididos em diferentes indústrias, sendo elas a indústria dos serviços financeiros e a indústria de produtos, serviços e recursos. Dentro de cada uma destas indústrias temos ainda três áreas de trabalho: tecnologia, estratégias e operações e recursos humanos.

Este estágio foi desenvolvido na área de consultoria, na indústria de serviços financeiros, dentro do grupo de estratégias e operações. No entanto, nem sempre os projetos trabalham apenas com um grupo, sendo por vezes necessário combinar diferentes grupos de trabalho de forma a obter a melhor solução.

Uma das grandes vantagens competitivas da Deloitte é a grande quantidade de informação da qual esta dispõe. Todas as empresas dos vários países cruzam informação e fazem uso de todo o conhecimento global, sendo praticamente nulos os projetos que têm de ser desenvolvidos

de raiz. Por norma o método de trabalho passa por verificar casos similares e tentar aplicar soluções semelhantes, testando as várias soluções até se traçar o caminho acertado.

A Deloitte tenta ainda criar equipas multidisciplinares contratando pessoas das mais diversas formações. Embora tenha uma grande componente de profissionais formados nas áreas de gestão, finanças e contabilidade aposta também em física, matemática e diversas engenharias. Procura-se uma formação com uma forte componente de tecnológica mas existe lugar também para outras engenharias como civil ou biomédica. Com esta estratégia é espectável que existam mais ideias e mais diversificadas, tendo em conta a vasta gama de conhecimentos.

A aposta mais recente da empresa e a que contém a mais variada formação de profissionais é a Deloitte Digital. Esta nova área de trabalho, que desenvolve projetos em todas as indústrias, pretende romper com todas as barreiras e diversificar o mercado. Tendo em conta a rápida evolução tecnológica que se está a verificar atualmente, a Deloitte entendeu que teria de aumentar o seu conhecimento. Esta componente aborda áreas como robótica e inteligência artificial e pretende desenvolver o conceito de consultoria digital e criar produtos e soluções mais inovadores de forma a crescer com os seus clientes.

1.1.2 A carreira

No esquema presente na figura 1.2 é possível observar os vários cargos que podem ser ocupados na Deloitte. Sendo inicialmente analista, um profissional pode subir hierarquicamente até atingir o nível de sócio.

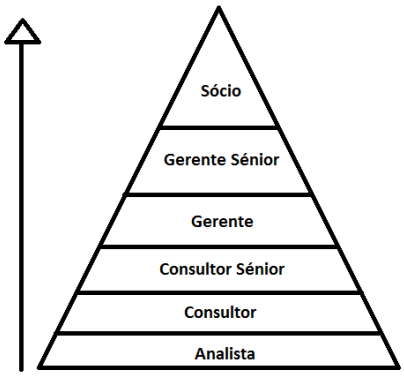


Figura 1.2: Esquema representativo da hierarquia de funções presentes na Deloitte.

Tendo em conta que existem mais analistas do que consultores, tal como podemos verificar pela figura 1.2, é necessário então definir a evolução da carreira. O principal fator neste processo é a produtividade de cada profissional e como tal a avaliação que este tem por parte dos seus superiores.

É necessário perceber o que é um bom consultor. Por um lado é favorável ter uma habilidade natural para ser consultor, sendo que por norma este gosta de resolver problemas para os seus clientes, sendo uma pessoa idealmente muito voltada para a inovação.

Existem no entanto outros fatores que auxiliam neste trabalho:

- Capacidade para perceber como funciona uma organização como um todo, assim como perceber todos os seus componentes;
- Aptidão para se relacionar com as pessoas de todos os níveis da organização;
- Confiança para mostrar uma ideia em que se acredita e a humildade de admitir um erro;
- Competência de sintetizar uma grande quantidade de dados em informação efetiva para ser apresentada num curto período de tempo;

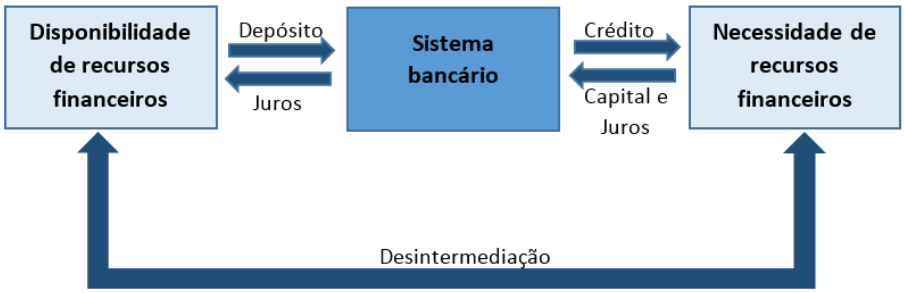


Figura 1.3: Esquema representativo do sector financeiro.

- Capacidade de mudança.

Podemos inferir que quanto mais uma pessoa se identifica com cada um destes pontos e como tal apresenta um maior rendimento mais eficaz é a sua evolução de carreira.

1.1.3 Sector Financeiro

O sector financeiro tem a principal função de armazenar recursos que são disponibilizados pelas poupanças das pessoas e transferi-los para o sector produtivo ou de consumo. Na figura 1.3 é possível verificar o esquema representativo deste processo. O sector financeiro é o elo de ligação entre entidades ou pessoas que querem investir recursos financeiros e aqueles que necessitam dos mesmos. A falta do sector financeiro levaria à desintermediação dos dois intervenientes.

Para além desta função, o sector financeiro também representa uma grande fonte de gestão de informação, sendo muito importante por exemplo em casos de fraudes.

Para além das instituições financeiras fazem parte também os mercados e órgãos reguladores.

A banca tem vindo a sofrer diversas alterações, nomeadamente em Portugal. É importante observar as mudanças ocorridas na sequência de privatizações que fizeram com que o setor necessitasse de processos de reorganização e de modernização. Partindo deste facto percebe-se que uma reorganização de estrutura implica ajustar o número de funcionários, assim como o número de balcões.

Por outro lado, esta indústria tem sofrido ainda alterações devido a fatores externos. Pode-se destacar a inovação tecnológica sem precedentes. Na realidade atual existe uma tendência a realizar tarefas através da internet, sendo a disposição dos clientes para uma possível deslocação a um balcão cada vez menor. Esta realidade aumenta a tendência de fecho de balcões.

Um dos grandes desafios da consultoria é ajudar as instituições financeiras a adaptarem-se a esta nova realidade, aproximando-os do consumidor. A cada dia que passa, a exigência do mercado aumenta sendo necessário aumentar a proximidade com o cliente. A Deloitte Digital serve precisamente para ajudar as empresas nestas questões, sendo uma aposta recente da empresa, e que mostra a necessidade que todas as empresas têm de inovar, até mesmo a própria Deloitte.

Este estágio foi desenvolvido com o intuito de criar um método facilitador na tomada de

decisões por parte dos bancos, no que diz respeito aos seus balcões.

1.2 Sistemas de Informação geográfica (SIG)

Existem informações que podem ser mais facilmente interpretadas através de uma componente visual de representação. Se pensarmos numa rede de instituições bancárias, de imediato tentamos visualizar um mapa das mesmas. Assim sendo, é importante poder perceber a nível geográfico, as envolventes de diferentes balcões, de forma a tomar decisões sobre os mesmos.

Nas sociedades modernas a localização das populações e as suas movimentações diárias tornaram-se muito mais complexas e fluidas no tempo. Onde há vinte anos existia uma região comercialmente florescente hoje pode ser uma zona vazia e abandonada. Na origem destes processos podem estar variações na rede de transportes ou deslocalizações de grandes centros agregadores. Esta complexidade dificulta o processo de tomada de decisão sobre a melhor localização para cada uma das agências. Gerir e otimizar racionalmente uma rede com centenas de balcões em circunstâncias tão complexas requer a obtenção e cruzamento de imensa informação. Por este motivo um SIG irá funcionar de forma a estabelecer pontos-chave na avaliação da rede.

Para além da densidade populacional e das vias de transporte existem inúmeros tipos de informações geográficas muito úteis para um negócio. Fatores como a distribuição etária da população, nível literário, atividade profissional, concorrência, etc, poderão ser muito úteis. Todas estas variáveis poderão ser cruzadas e tratadas estatisticamente e por norma representadas em mapas e gráficos.

Consoante os objetivos pretendidos é necessário ter em atenção a granularidade do estudo. Em alguns casos a escala do distrito pode ser suficiente, noutros basta a freguesia, frequentemente o ideal é a escala de um troço de rua ou de uma praça. Em alguns negócios a escala pode tornar-se extremamente fina: num hipermercado a escala do corredor ou da prateleira é crucial e atualmente é fortemente monitorizada sob diversas formas. A localização da exposição de um produto determina uma parte relevante das suas vendas e tem um preço para o fabricante.

Tendo em conta toda esta dinâmica e complexidade, é necessário criar meios para que se possa facilmente mapear a distribuição de pessoas, bens ou produtos. Este processo pode ser efetuado através de um SIG. Assim sendo é necessário explorar este conceito.

O termo GIS (SIG em português) veio simbolizar uma tecnologia, uma indústria, uma forma de fazer as coisas. O nome "Geographic Information Systems" foi utilizado pela primeira vez por Tomlinson em 1966.

No passado, a recolha, tratamento e armazenamento de informação era uma função bastante difícil e complexa. O aparecimento dos computadores e das tecnologias de informação e comunicação facilitaram esta tarefa. A evolução tecnológica permitiu o desenvolvimento dos SIG da forma como se apresentam atualmente.

Um SIG pode ser definido como um sistema de hardware, *software*, informação espacial, procedimentos computacionais e recursos humanos que permite e facilita a análise, gestão ou representação do espaço e dos fenómenos que nele ocorrem. Normalmente, este sistema

envolve conhecimentos multidisciplinares que conciliam áreas tais como, Geografia, Cartografia, Ciência da Computação, Sensorização Remota, Levantamento de Campo, Estatística, Matemática, Engenharia.

Tendo em conta a complexidade e a subjetividade envolvida na definição de um termo como este, analisaram-se várias definições. Neste documento apresenta-se a definição elaborada em 1990 por Marble baseada em quatro subsistemas detalhados [4]:

1. **Recolha:** Uma base de dados de entrada que recolhe e/ou processa dados espaciais provenientes de mapas, sensores, etc.

2. **Armazenamento:** Um subsistema de armazenamento de dados que organiza os dados espaciais recolhidos numa forma que permita que os mesmos sejam rapidamente recuperados pelo utilizador para uma subsequente análise, assim como rápidas e eficazes alterações e correções nos dados da base de dados.

3. **Processamento:** A manipulação dos dados e análise que compreende várias tarefas tais como a alteração da forma dos dados ou a produção de estimativas de parâmetros, otimização e simulação de modelos.

4. **Representação:** Um subsistema que descreve todos os dados originais ou parte deles assim como a manipulação dos dados e o output na forma de tabelas ou mapas. A criação deste mapa envolve a cartografia digital. Esta é uma área que representa uma expansão conceptual da cartografia tradicional e uma mudança de ferramentas para criar os mapas cartográficos.

Podemos concluir que a definição para o SIG é flexível e que depende não só da perspetiva de cada um como do objetivo pretendido.

Para a criação de um SIG existe uma série de tomadas de decisão a efetuar. Os passos podem ser descritos por:

1. Definir quais são os dados em que estamos interessados para a construção do nosso sistema.

2. Verificar se é possível obter os mesmos e, no caso de ser possível, de que forma conseguimos obtê-los, analisando também se estes implicam custos. Nesse caso é necessário avaliar a relação custo/benefício.

3. Processar os dados e avaliar a informação contida. Este passo compreende a criação de gráficos e mapas que melhorem a perceção dos resultados, sendo necessário na criação dos mapas definir a projeção a utilizar e combinar diferentes informações através de camadas.

Nas duas próximas subsecções vamos explorar os pontos 2 e 3 deste processo.

1.2.1 Recolha de Informação

Tal como foi referido no início da secção 1.2 existem vários dados geográficos a ter em conta na criação de um SIG, sendo que esta recolha de informação tem de ser efetuada com base numa análise prévia dos dados relevantes.

Uma fonte oficial de informação é o Instituto Nacional de Estatística (INE) que disponibiliza uma grande quantidade de dados. A informação mais atual está disponível no Censos 2011 no qual existem vários dados catalogados geograficamente e com vários graus de granularidade. Da informação recolhida nos censos podemos obter dados sobre a densidade populacional em Portugal bem como informações importantes sobre a qualidade de vida das pessoas tais como o número de casas que contêm saneamento ou o número de pessoas que reside em cada habitação.

Existe uma grande quantidade de informação útil disponível *online*. Por exemplo, através do *site* de um banco é possível aceder à morada de cada um dos balcões mediante uma pesquisa. A repetição exaustiva deste tipo de tarefas é penosa e demorada quanto efetuada por um operador, contudo em muitos casos é possível automatizar a tarefa com recurso ao desenvolvimento de código em linguagens de programação. Esta tarefa denomina-se por *data mining*. Na figura 1.4 está presente um exemplo ilustrativo de uma destas tarefas, sendo o código utilizado neste caso para extrair informação sobre os centros de saúde. Através deste algoritmo é possível procurar todas as unidades de saúde, alterando o número que constitui o *link* da figura, nomeadamente a linha 13. Esta procura permite aceder ao código HTML de cada uma das páginas que contem unidades de saúde. De seguida é necessário verificar para cada unidade os caracteres em que se encontra a morada e as respetivas posições. Por fim recolhe-se a morada correspondente a cada uma delas. Uma das linguagens adequada para esta tarefa e a utilizada no exemplo é o Python.

Podemos ainda pedir os dados a alguma instituição; nesse caso é necessário verificar se estas informações têm algum custo associado e decidir se o custo se justifica tendo em conta a importância dos dados.

1.2.1.1 Python

Python é uma linguagem de programação de alto nível concebida no final de 1989, por Guido van Rossum no Instituto de Pesquisa Nacional para Matemática e Ciência da Computação, nos Países Baixos.

Esta linguagem é de fonte livre o que faz com que haja uma ampla comunidade de utilizadores. Por este motivo existe uma vasta gama de bibliotecas que tornam a programação mais fácil, rápida e com um grande suporte *online*.

O Python tem bibliotecas que possibilitam o acesso ao código HTML das páginas de interesse. Depois de obtido o código é necessário encontrar um padrão que permita uma recolha sistematizada dos dados necessários através da pesquisa de um determinado conjunto de caracteres que se encontram exatamente antes ou depois da informação pretendida. A título de exemplo mostra-se na figura 1.4 o código utilizado para recolha das moradas dos centros de saúde.

```
def localizacao(ini,fin):
    textfile_txt=open("C:Users/Centrosloc.txt",'w')
    local=[]
    x=range(ini,fin)
    for i in x:
        site="http://www.portaldasaude.pt/Portal/servicos/prestadoresV2/?providerid="
        htmlfile=urllib.urlopen(site+str(i))
        htmltext=htmlfile.read()
        if 'lblAddress">' in htmltext:
            if '"item_head">\r\n                Centro de Sa' in htmltext:
                posicaoM= htmltext.find('lblAddress"')
                Morada=htmltext[posicaoM+12:posicaoM+70]
                morada=Morada.split("</span")[0]
                morada_final=codecs.decode(morada,'utf-8')
                posicaoF= htmltext.find('PostalCodeDescription')
                Freguesia=htmltext[posicaoF+23:posicaoF+80]
                freguesia=Freguesia.split("</span")[0]
                freguesia_final=codecs.decode(freguesia,'utf-8')
                local.append( morada_final+", "+freguesia_final)
                textfile_txt.write(morada_final+", "+freguesia_final+'\n')
            else:
                print "Hospital"
        else:
            print "Nao existe"

    textfile_txt.close()
```

ciclo for altera o link da página que procuramos
#através do range x defenimos a lista de números a usar
abre o código html
#leitura do código html da página
#os caracteres 'lblAddress">' indicam se existe a página e antecedem a morada

#caracteres "</span" estão sempre no final da morada
#descodifica o texto
#caracteres 'PostalCodeDescription' antecedem a freguesia

Figura 1.4: Excerto de um código de Python para a leitura do código HTML de uma página do portal da saúde, e recolha das moradas dos centros de saúde com o auxílio de código em Python.

Por fim a escolha da linguagem a usar prendeu-se também com o facto de esta ser a linguagem de programação lecionada durante o curso e ser a que me encontro mais à vontade para desenvolver.

1.2.1.2 Qualidade dos dados

A robustez dos resultados a apresentar depende fortemente da qualidade de toda a informação recolhida, nos processos anteriormente descritos.

Após a recolha deve ser efetuada uma análise da fiabilidade dos dados através da procura de erros pontuais ou sistemáticos que alterem significativamente a qualidade dos mesmos. Possíveis falhas podem ser causadas pelo processo de recolha da informação mas sobretudo pelos erros intrínsecos dos dados originalmente recolhidos.

Quando processamos os dados podemos introduzir erros indesejados. Erros podem ainda ser causados pela identificação errada do sistema de coordenadas em que os dados se encontram ou por cálculos efetuados durante o tratamento estatístico.

No entanto o processo de identificação de erros nem sempre é efetuado com o intuito de os eliminar podendo essa informação ser usada para perceber como gerir essas diferenças da melhor forma. É assim muito importante compreender e garantir a veracidade dos dados recolhidos por forma a assegurar uma apresentação mais robusta dos resultados que pretendemos alcançar.

1.2.2 Tratamento e Representação da Informação

A forma como se apresenta a informação é decisiva pois permite ao utilizador ter uma melhor perspetiva dos resultados. Quando temos uma grande quantidade de dados é crucial existir um tratamento prévio para que produzam informação útil. Os resultados são por norma sumariados em tabelas e gráficos; no entanto, a informação geográfica ganha outra dimensão quando representada por mapas topográficos.

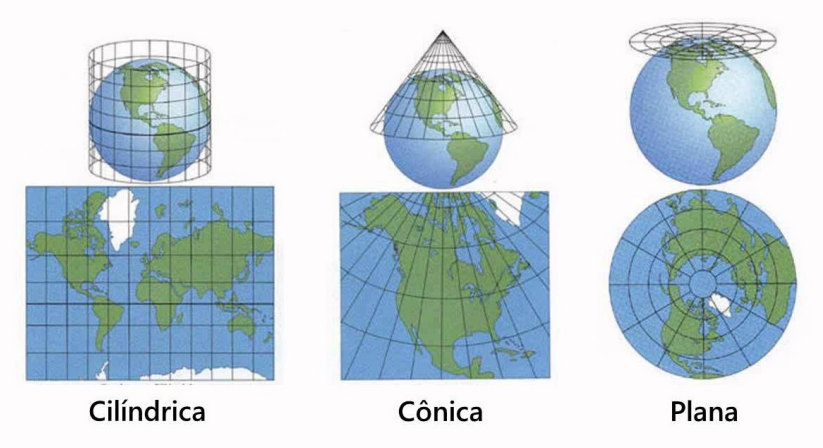


Figura 1.5: Principais projeções da Terra. Destacam-se a cilíndrica, cónica e plana.

A representação da superfície curva da terra num mapa com uma superfície plana implica inevitáveis distorções. Para resolver este problema temos de efetuar a projeção da superfície terrestre, sendo este um método matemático, pelo qual a superfície curva da terra é representada sobre uma superfície plana. Existem diferentes projeções; para exemplificar representamos na figura 1.5 os três tipos principais: cilíndrica, cónica e plana. Verifica-se que em qualquer projeção existe sempre alguma zona do globo distorcida: na projeção plana a parte externa da representação (topo do mapa, sendo definido de acordo com o local que se pretende representar qual a parte do globo na zona central) apresenta-se mais "esticada". Para evitar distorções, utiliza-se a projeção que considere o local a representar o mais central possível.

Figura 1.6: Representação da mesma morada em diferentes sistemas de coordenadas geográficas.

Para além das projeções, também o sistema de coordenadas é importante para a definição dos pontos no nosso mapa. Os sistemas de coordenadas podem ser definidos através de coordenadas tridimensionais ou bidimensionais. Os sistemas de coordenadas cartesianas bidimensionais, que são os utilizados quando representamos um mapa de forma plana são obtidos a partir das coordenadas elipsoidais, por intermédio de uma projeção cartográfica.

Existem muitos sistemas de coordenadas, que se adaptam de melhor ou pior forma ao local que queremos representar. A título de exemplo, na figura 1.6, encontra-se representada uma mesma localização no mapa de Portugal em dois sistemas de coordenadas diferentes. A projeção WGS84 é usada pelo Google Maps e pelo sistema de GPS, contudo esta não é a projeção na qual o mapa de Portugal está menos deformado.

A representação ideal para Portugal é a ETRS89. Por este motivo os Censos 2011 fornecem

todos os dados neste sistema de coordenadas. Da figura 1.6 concluímos que a confusão no sistema de coordenadas utilizado pode originar diferenças bastante significativas. Utilizam-se as coordenadas obtidas no sistema WGS84 para representar mapa de Portugal e um ponto na cidade da Póvoa de Varzim, distrito do Porto e representamos essas mesmas coordenadas nos dois sistemas de projeção (WGS84 e ETRS89). Como podemos verificar os dois sistemas de coordenadas não coincidem.

Na criação de mapas em *softwares* GIS o processo é feito através da sobreposição de camadas distintas. Com camada entende-se um mapa que contém informação útil para o nosso estudo, sendo por exemplo o mapa dos rios de Portugal uma camada que só contém exclusivamente os rios, desenhados como linhas e que pode estar sobreposto a um mapa de polígonos que representam os distritos de Portugal. A título de exemplo podemos imaginar que pretendemos criar um mapa com a densidade populacional e o número de estradas em Portugal então criamos a sobreposição de 3 camadas: uma com o mapa de Portugal, outra com a densidade populacional e por fim uma com as estradas portuguesas.

Para representar os valores no mapa existem duas abordagens principais:

- O modelo matricial, no qual a área é dividida numa rede regular e a posição de cada célula é definida pela linha e coluna onde se encontra sendo a cada uma associada um valor.

- O modelo vetorial, que pode ser representado por:
  - Pontos que servem para representar coordenadas.
  - Linhas que para além de construtores para os polígonos também podem representar entidades lineares como rios ou estradas.
  - Polígonos que representam entidades do tipo área.

Na figura 1.7 podemos ver a sobreposição de uma camada vetorial de polígonos, em que cada polígono representa um concelho de Portugal continental, com uma camada vetorial de pontos, na qual cada ponto representa um centro de saúde. Criamos assim mapa que nos permite avaliar o número de centros de saúde por concelho. Podiam ainda ser acrescentadas informações sobre, por exemplo, a população presente em cada concelho, dos transportes presentes, das faixas etárias da população, entre vários outros dados.

Este tipo de *software* permite então efetuar um estudo geográfico de elevado interesse, uma vez que possibilita a análise de uma grande quantidade de informação que pode ser visualizada em conjunto. Mais ainda, esta informação pode ser trabalhada em vários níveis de granularidade. O censos apresenta uma grande fonte de informação, fornece mapa com níveis de granularidade que vão desde a subsecção ao distrito. Para que exista uma noção do que representa uma subsecção, que são áreas instituídas pelo instituto nacional de estatística, estes apresentam as freguesias divididas em áreas às quais chamam de secção e cada uma destas está dividida em mais áreas que são as subsecções. Estas áreas não apresentam dimensão fixa, dependendo se estamos a trabalhar no centro urbano ou num meio rural.

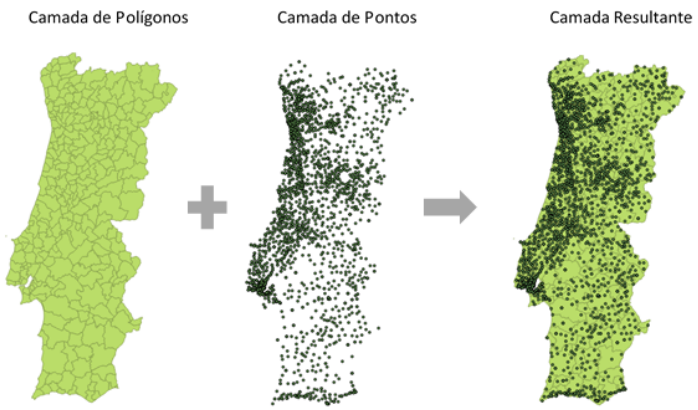


Figura 1.7: Representação da sobreposição de duas camadas vetoriais. Uma camada de polígonos associada a uma camada de pontos origina outra camada com as duas informações.

1.2.3 *Software* de GIS

Um *software* de GIS (geographic information system) permite a visualização, edição e análise de dados georreferenciados. Podem ser utilizados para a criação de mapas com várias camadas usando diferentes projeções e disponibilizam uma vasta gama de ferramentas para tratamento e análise. Dos *softwares* comerciais, o mais referenciado é o ArcGIS. Contudo existem *softwares* de código de fonte livre como gvSIG e QGIS que são bastante completos.

No decorrer deste projeto será utilizado o QGIS pois permite a execução de todas as tarefas pretendidas, é bastante intuitivo, apresenta uma grande quantidade de bibliografia disponível e tem fóruns de discussão muito ativos. O QGIS, inicialmente conhecido como quantum GIS foi desenvolvido em 2002 por Gary Sherman e atualmente está associado à Open Source Geospatial Foundation. A versão 1.0 foi lançada em Janeiro de 2009. Podemos verificar na figura 1.7 um exemplo prático das aplicações de QGIS, no qual se combina informação de diferentes fontes, cada uma representada numa camada distinta e se combina a informação numa só camada. Neste caso foi efetuada a junção do mapa de Portugal por concelho (camada vetorial de polígonos) com os centros de saúde portugueses (camada vetorial de pontos).

1.3 O estágio

Os capítulos seguintes apresentam a componente prática desenvolvida durante este estágio na empresa.

No capítulo 2 apresentamos todo o processo de criação de um SIG, a partir de informação sobre clientes da qual constava um conjunto de informações como crédito e depósito associadas a uma morada devidamente encriptada. Neste capítulo é apresentado o método desenvolvido para a criação de um mapa ao qual conseguimos associar informação através de coordenadas geográficas. Este apresenta informação sobre a comparação de *strings*, as ferramentas necessárias para a criação do sistema, as fontes de informação utilizadas e por fim a explicação do algoritmo final.

No capítulo 3 são abordadas várias conclusões relacionadas com o documento inicialmente

fornecido para o trabalho. As conclusões estatísticas são obtidas apenas utilizando esse ficheiro e as conclusões a nível geográfico são obtidas partindo dessa informação mas utilizando o mapa descrito no capítulo 2.

Por fim existe um capítulo de conclusão de todo o trabalho realizado.

2 Criação de um mapa com moradas associadas

O problema presente prende-se com a criação de um mapa ao qual se pretende associar informação através de coordenadas geográficas e de moradas. Obtido este mapa é possível analisar a informação dos clientes, com a finalidade de chegar a um estudo geográfico de mercado. Como ponto de partida temos um conjunto de informações para cada cliente: crédito, depósito, idade, habilitações e morada devidamente codificada.

É necessário criar uma forma de analisar os dados ao nível geográfico. Para esta tarefa utilizam-se *shapefiles* que apresentam a informação através de mapas. Estes mapas podem ser manipulados através de *software* como por exemplo o QGIS e apresentam informação associada.

Na figura 2.1 está representado um esquema da forma como se articulam as fontes de informação usadas para a criação do mapa pretendido. Inicialmente recorre-se a uma *shapefile* do Open Street Maps (OSM) que contém todas as estradas do território nacional. Esta camada contém associada a cada rua informações quanto à morada e ao tipo de via. Recorre-se ao OSM pois é uma fonte de informação livre que pode ser utilizada como ponto de partida para a criação da base de dados de informação geográfica.

Dada a elevada dimensão da informação é útil reduzir as zonas de estudo de forma a tornar os processos a realizar mais rápidos. O mapa disponibilizado pelo OSM contém as estradas associadas a todo o país. Pode-se restringir o mapa às estradas de um distrito, cidade ou freguesia. Para diminuir a área a analisar é necessário recorrer aos mapas do INE.

O INE contém um conjunto de dados de acesso livre fazendo parte destes toda a informação recolhida através dos censos. Os dados obtidos pelos censos têm como principal objetivo recolher, agrupar e publicar informações demográficas, sociais e económicas sobre uma população num dado momento. Este conteúdo está disponível num formato *csv (comma separated value)* e pode ser associado a uma camada representativa de Portugal, no formato *shapefile*. Através do INE é possível obter um mapa de Portugal como um todo, um mapa da região norte, centro ou sul ou um mapa das cidades portuguesas de forma isolada, ou seja, posso obter por exemplo apenas o mapa da cidade do Porto. Os mapas de cada localização são usados para limitar as estradas que pretendemos estudar. A interação entre o OSM e o INE está representada esquematicamente na figura 2.1.

O primeiro problema deste projeto está relacionado com as falhas na informação do OSM. No mapa do OSM as estradas são representadas por partes de linhas denominadas por troços. Os troços representam estradas, no entanto nem todas têm um nome associado. Para resolver este problema é necessário obter informação de outras fontes, neste caso recorre-se ao GAPI[1]

[1]GAPI refere-se à aplicação disponível para obter informação do Google Maps

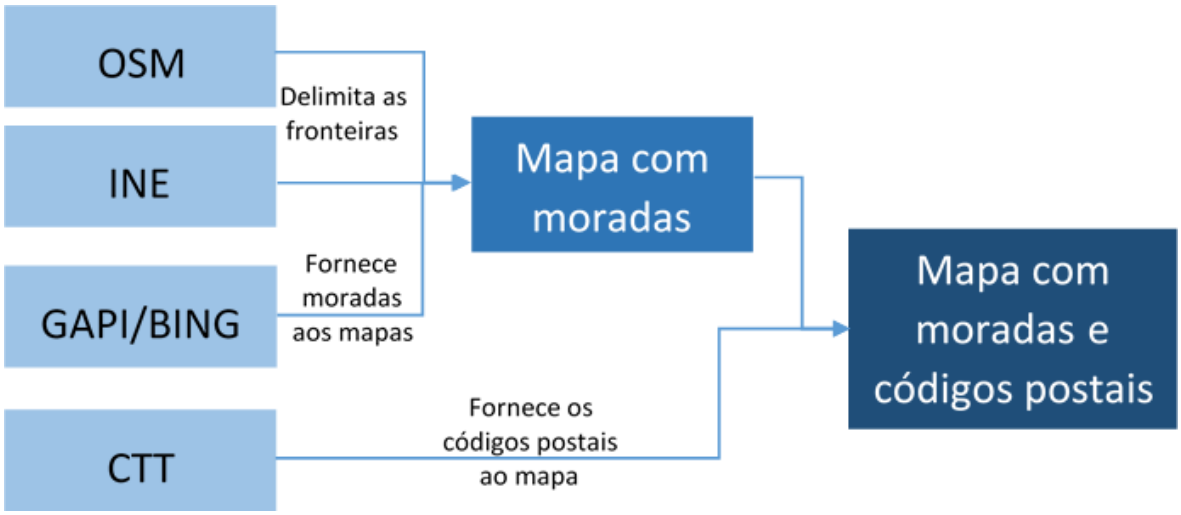


Figura 2.1: Esquema da criação do mapa com as moradas associadas.

e ao BING[2]. O *software* de informação geográfica permite obter a coordenada geográfica do ponto médio de cada troço. Fornecendo ao GAPI e ao BING esta coordenada geográfica é possível obter as respetivas moradas. O conhecimento da morada é crucial para podermos localizar geograficamente cada cliente.

Associar dois nomes de ruas é um processo bastante complexo pois as ruas podem estar escritas com diferentes abreviaturas, com erros nas preposições ou erros de escrita. A forma mais exata de comparar duas localizações é recorrendo ao código postal, por este motivo sempre que o cliente tiver uma morada com código postal irá dar-se preferência a esta informação preterindo o nome da rua. Para esta tarefa está disponível uma base de dados dos CTT com o conjunto de todos os códigos postais de sete dígitos associados a um nome de rua e freguesia. No esquema representado na figura 2.1 evidenciam-se as interações entre as diferentes fontes de informação. O resultado final de troços com moradas e códigos postais é o produto da combinação dos diferentes elementos fornecidos pelas quarto fontes. A melhor fonte para o código postal são os CTT mas para associarmos estes às coordenadas geográficas precisamos da interação do GAPI e do BING.

No decorrer desta fase pretende-se encontrar um mapa no qual a cada troço tenha filiado uma morada (nome de rua, número da porta, freguesia e código postal) que possa ser trabalhado em QGIS e à qual possamos associar os clientes. Podemos definir as tarefas que se irão seguir nos seguintes tópicos:

1. Definir um mapa em que a cada rua tenha associada uma coordenada;
2. Procurar qual a morada associada a cada coordenada;
3. Associar a cada morada um código postal dos CTT;
4. Juntar as moradas, com código postal ao mapa;
5. Juntar a informação de cada cliente ao mapa através do código postal ou caso este não esteja disponível através da morada.

[2]BING refere-se à aplicação disponível para obter acesso à informação do Bing Maps

Todos estes processos são tratados nesta secção, que inclui a recolha, armazenamento e processamento da informação de um SIG.

2.1 Open Street Maps

2.1.1 Valências

O Open Street Maps é um projeto de mapeamento livre que tem como objetivo criar um mapa do mundo através de diferentes contribuições, num modelo de desenvolvimento semelhante ao do wikipédia. O projeto foi criado em Julho de 2004 por Steve Coast e para além das estradas fornece construções, lagos, rios, etc.

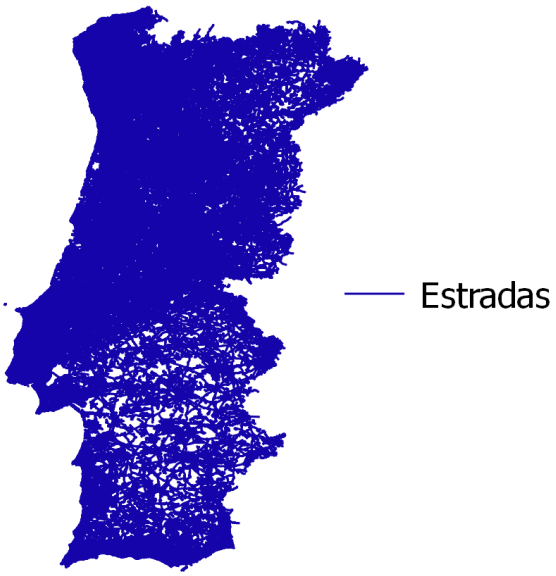


Figura 2.2: Gráfico recolhido no OSM, em formato *shapefile*. Este mapa é constituído apenas por linhas que representam as estradas de Portugal.

A sua principal vantagem face a outros sistemas semelhantes é o facto de este ser um *software* livre. Esta característica permite a qualquer utilizador criar o seu próprio mapa.

Utilizam-se estes mapas, no formato *shapefile*, formados por linhas que representam as estradas. Estes permitem-nos associar coordenadas geográficas a cada um dos troços bem como outras informações (e.g. nome de rua ou código postal). Na figura 2.2 pode ser visualizado o mapa com todas as estradas portuguesas, disponível a partir de uma *shapefile* do OSM [13].

Estas *shapefiles* contemplam diversos tipos de vias, classificados com diferentes nomes. Na tabela 2.1, estão descritos os tipos de ruas usados no mapa e a respetiva descrição de cada tipo.

Para simplificar o trabalho posterior, é importante eliminar da *shapefile* os acessos nos quais não é possível existir nenhuma morada. Tendo em conta a descrição de cada um dos tipos de acesso e analisando num mapa como por exemplo no Google Maps as mesmas ruas, percebe-se que não existe nenhuma habitação nos seguintes percursos: cycleway, bus_stop, elevator, footway, motorway, motorway_link, path, platform, raceway, service, services, track, trunk, trunk_link.

Partindo da *shapefile* inicial e excluindo os troços nos quais não é possível existir moradas obtém-se uma *shapefile* com todas as linhas das estradas portuguesas que podem conter habitações. Para limitar a zona de estudo a alguma área específica é necessário o auxílio de informação dos censos. Estes contêm mapas formados por polígonos em que cada um pode representar uma secção, freguesia ou cidade. Na figura 2.3 está representada a sobreposição das estradas do OSM com a área metropolitana do Porto obtida a partir do INE. Partindo

destas duas camadas é possível extrair o mapa que contém apenas as estradas do Porto, tal como representado na imagem.

Nome	Descrição
cicleway	Ciclovias.
bus-stop	Paragens de autocarros (ex.: Campanhã).
elevator	Teleférico.
living-street	Ruas com habitações (ex.: parques de estacionamento de bairros)
motorway	Autoestradas.
motor_link	Acesso às autoestradas.
path	Caminho de terra.
pedestrian	Passeios e ruas fechadas à circulação de veículos motorizados (ex.: rua da junqueira, travessa da lage).
platform	Cais, paragem (no caso do porto refere-se a paragem de autocarros perto do bom sucesso).
primary	Estradas residenciais.
primary_link	Estradas residenciais.
raceway	Jardins.
residential	Estradas residenciais.
road	Estradas residenciais.
secondary	Estradas residenciais.
secondary_link	Estradas residenciais.
service	Serviços (ex.: bombas de gasolina, parques privados)
services	Serviços (ex.: bombas de gasolina, parques privados)
steps	Escadas.
tertiary	Estradas residenciais.
tertiary_link	Estradas residenciais.
track	Caminhos de terra.
trunk	Autoestradas.
trunk_link	Autoestradas.
unclassified	Caminhos de terra e estradas com habitações.

Tabela 2.1: Classificação das estradas usada pelo OSM.

A cada uma das linhas do OSM está associada a informação:

- osm_id: número único associado a cada troço;
- nome da rua: nome da rua representada pelo respetivo troço;
- tipo de troço: tipo de via associada; esta classificação prende-se com os termos descritos na tabela 2.1.

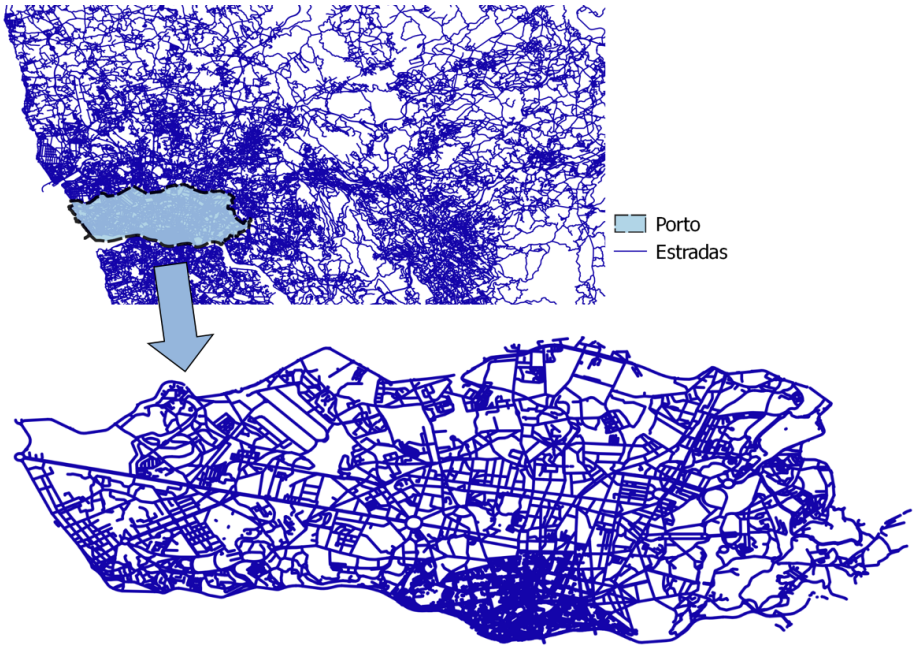
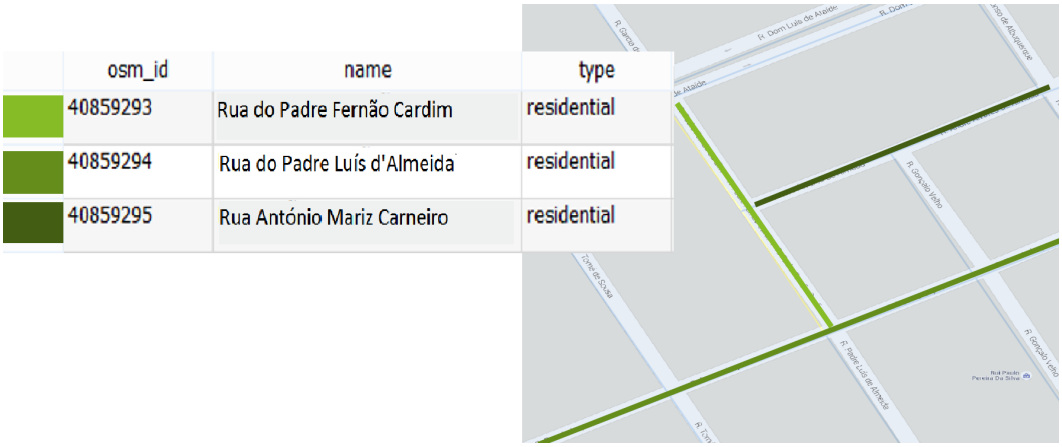


Figura 2.3: Partindo de um mapa de estradas e de um mapa de com a delimitação de uma dada zona é possível obter as estradas dessa zona apenas.

Na figura 2.4 podem ser observados três das linhas da *shapefile* do OSM e a tabela que contém a informação de interesse associada a cada uma. Os troços foram sobrepostos ao mapa do *Google Maps* para tornar mais clara a sua visualização.



osm_id	name	type
40859293	Rua do Padre Fernão Cardim	residential
40859294	Rua do Padre Luís d'Almeida	residential
40859295	Rua António Mariz Carneiro	residential

Figura 2.4: Esta figura representa três linhas do mapa das estradas de Portugal e um excerto da tabela associada a estas linhas. A imagem de fundo não faz parte da *shapefile* do OSM usada, sendo esta uma imagem do Google Maps, utilizada para facilitar a visualização da informação.

2.1.2 Dificuldades

Inicialmente quando foi pensada a utilização do OSM esperava-se que a cada troço estivesse associada uma morada. O primeiro problema que se torna evidente neste trabalho prende-se com o facto de não existir nome de rua para todos os troços, tal como pode ser ver observado através da tabela de informação presente na figura 2.5. Existem casos em que todos os campos

aparecem preenchidos, no entanto em outras situações o único campo preenchido é o osm_id.

Este facto ocorre por dois motivos:

- Falta de informação do OSM. Tendo em conta que esta é uma fonte criada por diferentes utilizadores que contribuem com informação, caso ninguém tenha acrescentado alguma rua esta irá estar em falta;
- Casos em que uma linha tem mais do que uma rua associada. É possível observar a rua com o osm_id 12544130, da figura 2.5, que é constituída por três ruas distintas mas que é representada apenas por um troço no OSM.

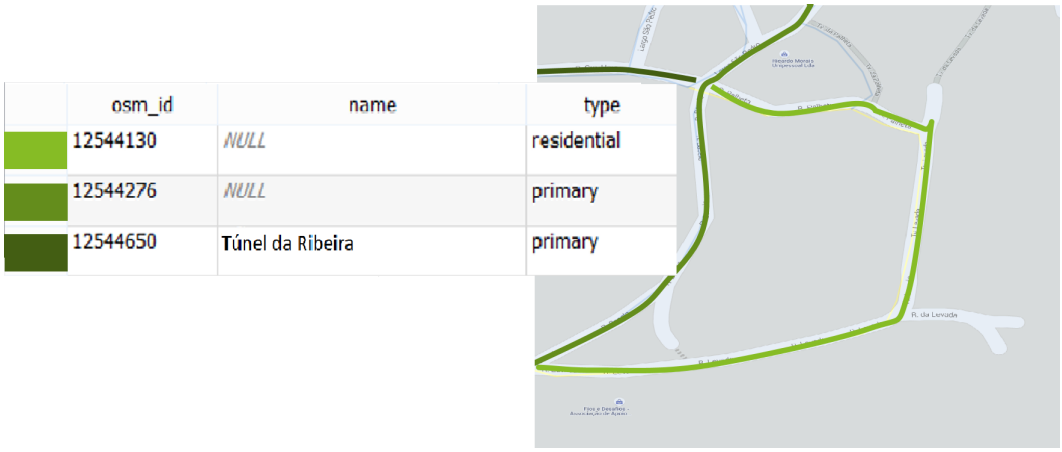


Figura 2.5: Este mapa contém um exemplo de linhas do mapa do OSM, nas quais existem linhas sem nome de rua associada.

Para a resolução dos casos nos quais os troços estão associados a mais do que uma rua, é necessário partir o troço.

Para realizar este processo uma possibilidade é definir o comprimento que pretendemos obter entre quebras. No entanto esse processo pode em alguns casos resultar numa grande quantidade de troços. Este processo dá origem à existência de troços que poderão conter ruas adjacentes.

Optamos por efetuar esta divisão através da quebra de todos os troços sempre que encontram intersecções com outras linhas. A solução de partir a cada cruzamento é útil pois consegue-se dessa forma tornar os troços pertencentes apenas a uma rua, sendo por norma a mudança de nome de rua associada a um cruzamento ou entroncamento. Esta opção do QGIS permite uma solução eficaz mas sem se obter uma quantidade exagerada de novos troços. Temos então que uma rua pode ser formada por um conjunto de troços, no entanto um troço só poderá conter uma única rua.

Na figura 2.6 podemos verificar que o mesmo troço da figura 2.5 (osm_id = 12544130) apresenta agora três troços diferentes. A cada um dos troços é possível associar um ponto central, ao qual chamamos centróide e este ponto tem associado a si um conjunto de coordenadas geográficas. Obtém-se assim um conjunto de troços, aos quais pertence apenas uma rua e que contêm associados a si um conjunto de coordenadas geográficas e em alguns casos um nome de rua.

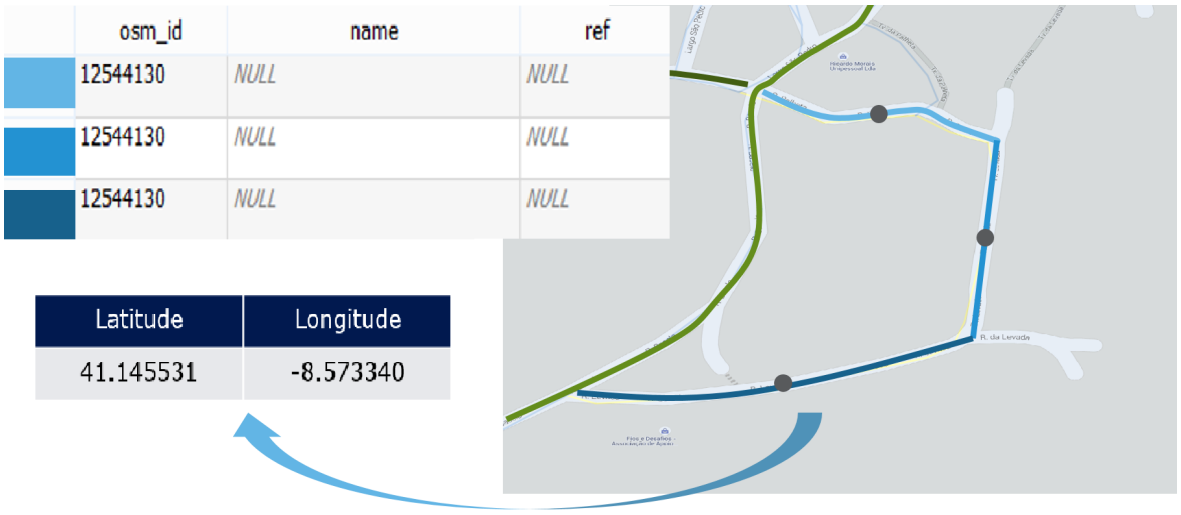


Figura 2.6: Gráfico com as linhas do OSM depois de partidas a cada entroncamento. Podemos verificar que a cada linha é possível associar o centróide (círculos representados na figura) e a cada centróide corresponde um conjunto de coordenadas geográficas, que neste caso está no sistema WGS84.

Para os casos em que o OSM não apresenta um nome de rua a si associado, é necessário obter outras fontes de informação para associar a cada troço uma morada. De seguida são descritas as outras fontes de informação utilizadas, bem como o método utilizado para associar os dados existentes.

2.2 Construção da base de dados

Para a realização deste trabalho é construída uma base de dados em que cada troço tem inicialmente associado a si um id único e a morada do OSM. Posteriormente são associadas as moradas do GAPI e do BING a cada um dos troços. Esta base de dados terá de conter ainda o número da porta e a freguesia correspondente a cada troço e por fim o código postal dos CTT correspondente. Nesta base de dados toda a informação é utilizada sem acentos, tendo em conta que para além de serem uma fonte de diferença entre as moradas, estes provocam dificuldade ao desenvolver os algoritmos necessários.

Tendo em conta que estamos a tratar de uma grande quantidade de informação é necessário que esta se mantenha organizada e que seja excluída aquela que não é relevante para o trabalho ou contém erros associados a si. Nesta secção irão ser discutidas as novas fontes de informação bem como os processos de tratamento utilizados.

2.2.1 Recolha de Informação - GAPI e BING

Tendo em conta que nem todos os troços representados no OSM apresentam um nome de rua associado, foi necessário procurar fontes alternativas. Esta análise, assim como todas as referidas neste capítulo, têm como base os resultados obtidos para o Porto.

Partindo do ficheiro que contém as coordenadas geográficas de todos os centróides dos troços é possível fazer uma pesquisa automática para obter as moradas correspondentes. A pesquisa das moradas é efetuada através de um determinado *link* que fornece o acesso a uma página

```
"formatted_address" : "R. Levada 400, 4300 Porto, Portugal",
"geometry" : {
   "location" : {
      "lat" : 41.1455075,
      "lng" : -8.57363
```


Figura 2.7: Excerto da página obtida através do GAPI. A morada recolhida é da forma "rua e número da porta, código postal, país".


```
▼<Location>
   <Name>Rua Levada 400, 4300-295 Porto, Portugal</Name>
  ▼<Point>
     <Latitude>41.14546</Latitude>
     <Longitude>-8.57362</Longitude>
   </Point>
```


Figura 2.8: Texto obtido através da pesquisa efetuada pelo BING. Podemos ver a sublinhado o texto obtido através da pesquisa automática.

escrita em formato *json*, tal como mostra o excerto da figura 2.7. Esta página contém várias informações como rua, número da porta, código postal, freguesia, cidade e as coordenadas geográficas associadas. Partindo então das coordenadas latitude e longitude, procura-se a rua que se encontra sublinhada na imagem 2.7. O *link* utilizado para as pesquisas associadas ao GAPI é

$$\text{“}https://maps.googleapis.com/maps/api/geocode/json?latlng=\text{”} + latitude, longitude$$

no qual se substitui as variáveis *latitude* e *longitude* pelos respetivos valores.

O *Google Maps* constitui uma fonte viável, a partir da qual é possível obter moradas para cada um dos troços e permite aceder à informação necessária, de forma relativamente simples. Contudo existe uma enorme desvantagem deste *software* que é apenas ser possível efetuar 2500 utilizações diárias de forma gratuita.

Depois de recolhida a informação e através de uma breve análise visual à qualidade dos dados é possível verificar que nem todos os casos que contêm duas moradas apresentam o mesmo resultado nas duas fontes.

Para resolver esta diferença, assumindo que não se sabe qual a fonte correta, é conveniente recorrer a uma terceira fonte. Neste caso utiliza-se o *Bing Maps* para verificar a qualidade dos resultados. Neste caso o *link* é do tipo

$$\text{”}http://dev.virtualearth.net/REST/v1/Locations/\text{”} + latitude, longitude + \text{“}o = xml\text{”}$$

Na figura 2.8 pode-se verificar um exemplo de uma página obtida através desta pesquisa.

Construiu-se assim a informação associada a cada um dos troços. Note-se que para alguns casos chegamos a ter três moradas para um único troço.

Durante o processo de recolha foram criadas 3 tabelas com a respetiva informação de cada fonte e que contêm os seguintes campos:

1. OSM: Rua; osm_id; ID

OSM	GAPI	BING
41.4%	100%	87.6%

(a) Percentagem de troços que contêm morada associada.

Três Fontes	Duas Fontes		Uma Fonte
	OSM-GAPI	BING-GAPI	
38.4%	3.1%	49.3%	9.2%

(b) Percentagem de troços em que temos três moradas associadas, duas ou apenas uma. São visíveis as fontes nos dois casos possíveis de se obter duas moradas.

Figura 2.9: Tabelas representativas dos dados existentes.

2. GAPI: Rua; número da porta; código postal e freguesia; osm_id; ID

3. BING: Rua; número da porta; código postal e freguesia; osm_id; ID

Na tabela 2.9a é apresentado o resumo do número de troços com morada para cada uma das fontes. O GAPI contém morada para 100% dos troços, seguido do BING que apresenta morada em 87.6% e o OSM que só apresenta morada em 41.4% dos casos.

Observando a tabela 2.9b é possível verificar que em 38.4% dos casos as três fontes apresentam moradas e apenas 9.2% do total apresenta apenas uma morada. Os restantes casos apresentam duas moradas, sendo possível identificar as percentagens correspondentes a ter morada do OSM e do GAPI e ainda casos em que apenas apresenta morada do BING e do GAPI.

Estas tabelas permitem ter uma ideia geral da quantidade de dados que estão disponíveis bem como da fonte mais completa.

Os nomes das ruas nem sempre coincidem, quer por erros ortográficos ou até mesmo por serem nomes de ruas distintos pelo que é necessário definir qual o nome correto para cada um dos troços. Os processos de organização e limpeza dos dados são descritos a seguir.

2.2.2 Limpeza dos dados

Nesta secção apresentam-se os procedimentos que podem ser realizados previamente com o intuito de facilitar a comparação entre fontes. Estão descritos processos que tratam a organização dos dados, a utilização de motores de pesquisa para excluir informação errada, a exclusão de ruas que não contêm habitações e a determinação de erros provenientes de diferentes resultados entre as fontes no caso de ruas próximas.

2.2.2.1 Organizar a tabela

Tendo em consideração que existem três fontes, uma das informações que se pode obter é a estimativa da fonte mais correta. Esta informação permite visualizar com mais clareza os dados e perceber qual a fonte que contém menos erros associados. Para isso é definido um grau de prioridade para cada uma das fontes.

Para efetuar este estudo consideram-se apenas os casos em que as três fontes contêm uma morada associada. Considera-se que uma morada está correta sempre que duas das fontes coincidem. Para pontuar cada uma das fontes, serão contabilizados o número de moradas corretas associada a cada fonte. Note-se que cada item só pode pontuar duas ou três fontes. Na tabela 2.2está representado um exemplo ilustrativo.

OSM	Contador OSM	GAPI	Contador GAPI	BING	Contador BING
ruadaagonia	1	ruadeagonia	1	ruadopadreborges	0
em502	1	ruamanuelmartins felgueiras	2	ruamanuelmartins felgueiras	1
estradanacional13	1	ruagomesdeamorim	3	ruagomesdeamorim	2
ruadamata	1	ruacidadedapovoa	3	ruadeagrovelho	2

Tabela 2.2: Tabela representativa do método utilizado para verificar qual a fonte mais correta. Vemos um contador cumulativo associado a cada uma das fontes ao qual é adicionado o valor 1 sempre que têm correspondência no mínimo entre duas fontes..

Se dividirmos o valor obtido no contador pelo total de troços analisados, obtemos a percentagem de certeza para cada uma das fontes. Os resultados obtidos foram: 85% para o GAPI, 73.7% para o BING e 72.7% para o OSM.

Tendo em conta estes resultados é possível organizar a tabela de forma a que a primeira morada seja sempre a mais provável e assim consecutivamente. Se existir morada do GAPI e do BING, ou do GAPI e do OSM o GAPI aparece em primeiro lugar. A exceção acontece para os casos em que existam moradas das três fontes mas a morada do BING seja igual à do OSM, nesse caso o BING é colocado em primeiro lugar.

Deste modo organiza-se a tabela com a fonte que é considerada mais correta em primeiro lugar e assim consecutivamente. Esta medida não resolve nenhum dos casos mas fornece uma ideia sobre os dados e a veracidade dos mesmos.

2.2.2.2 Otimizar o processo através de motores de pesquisa

Uma das formas de diminuir os erros iniciais é através de motores de pesquisa de páginas como o Google e o BING. Estes têm programas que percorrem de forma automática e periódica todas as páginas da internet e são capazes de detetar um conjunto de caracteres indicados por nós. Esta pesquisa retorna uma resposta que contém os caracteres pretendidos.

Quando se comparam ruas que apresentam nomes distintos constata-se que a maior parte dos casos estão relacionados com erros no nome da rua (ou seja o nome da rua não existe). Se for efetuada uma busca no motor de pesquisa da forma "nome da rua" "cidade", as ruas com nome errado não são encontradas. Mesmo para casos em que o nome da rua existe no mapa de uma determinada aplicação, esta no caso de não existir nao aparece no motor de pesquisa. A título de exemplo uma rua pode não ser encontrada pelo motor de pesquisa do Google e ser encontrada no Google Maps.

Nestes casos o Bing Maps ou o Google Maps contêm uma determinada rua que constitui um erro. Recorrendo ao motor de pesquisa para procurar essa rua associada à respetiva cidade este indica que não existe nenhum registo para a nossa pesquisa.

A forma de pesquisa que indica o nome da rua entre aspas seguido da freguesia também entre aspas implica que a pesquisa seja efetuada pelo nome da rua como um todo e não procure por

Figura 2.10: Exemplo de um caso no qual o motor de pesquisa do BING não encontra nenhum resultado.

exemplo apenas a palavra rua. Assim sendo, criou-se um processo automatizado que efetua uma pesquisa prévia no motor de busca, excluindo-se as moradas que não são encontradas. Na figura 2.10 pode ser observado um exemplo de pesquisa sem resultados através do BING. Verifica-se que quando não existe nenhuma morada encontrada pelo motor de busca, obtém-se a mensagem "Nenhum resultado encontrado", sendo esta a informação utilizada para verificar de forma automática se a morada existe. Esta pesquisa foi também efetuada no Ask que retorna a mensagem "did not match" quando não encontra nenhum resultado. Hoje em dia o mais provável é que todas as ruas estejam referidas em alguma página da internet.

Neste estudo não foi utilizado o motor de pesquisa do Google por não ter sido encontrada nenhuma solução para conseguir ler de forma automática a mensagem transmitida na página de pesquisa.

2.2.2.3 Excluir ruas que não contêm habitações

Tal como referido no capítulo 2.1.1, foram excluídos dos troços iniciais aqueles que representavam vias nas quais não existem habitações. No entanto, existem casos em que alguns tipos de troços, como as estradas nacionais apresentam partes nas quais não contêm habitações. As estradas nacionais por norma apresentam uma grande extensão ao longo da qual podem ter zonas com habitações e outras sem nenhuma morada.

A análise efetuada das fontes parece indicar que grande parte das regiões das estradas nacionais que contêm habitações apresentam outro nome associado.

A título de exemplo, podemos considerar o cado da "Estrada nacional 13", que está representada na figura 2.11 e que como se pode verificar contém uma determinada zona habitacional, assinalada a vermelho, na qual apresenta o nome "Rua Gomes de Amorim". Uma das formas

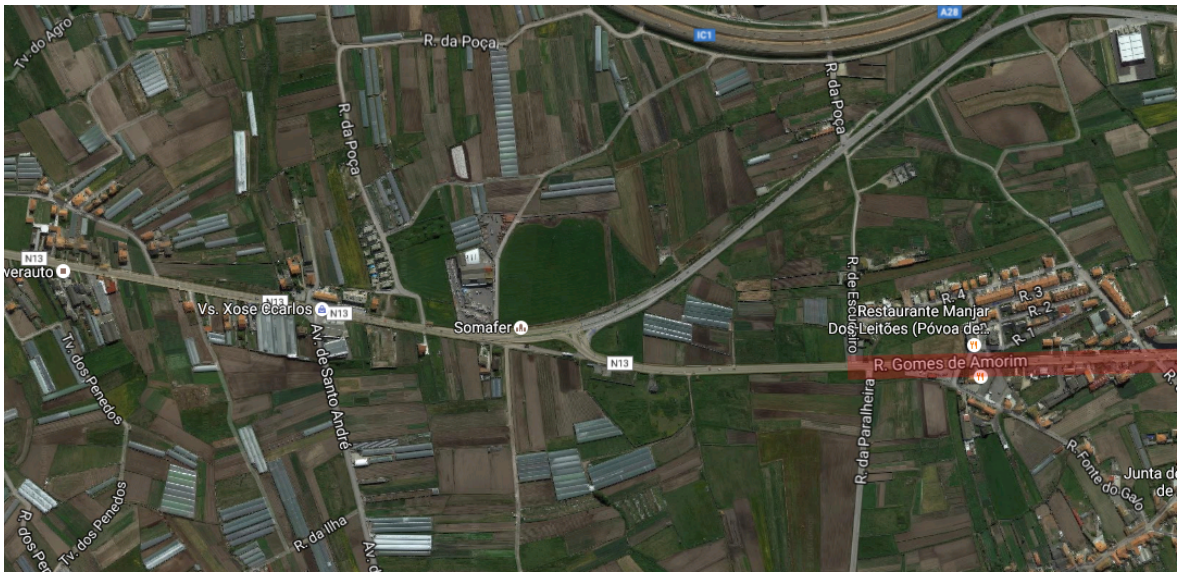


Figura 2.11: Representação de uma estrada nacional na Póvoa de Varzim.

de conseguir limpar os registos associados a estas estradas é assegurar que estão presentes na lista dos CTT, pois se existirem habitações em alguma zona da estrada nacional esta contém um código postal associado. Este é o caso que escolhemos para ilustrar na figura.

Com este método é possível resolver 1.2% dos troços.

2.2.2.4 Verificar a distância entre ruas distintas

Outra das limpezas que pode ser efetuada prende-se com a confusão que por vezes se verifica nos limites de uma rua. O principal problema consiste na identificação de ruas vizinhas pelas diferentes fontes. A figura 2.12 mostra um exemplo do que acontece nestes casos. Na primeira imagem, obtida através do Google Maps, verifica-se que a rua contida dentro do círculo tem o nome "Rua Viana do Lameiro". A segunda figura, que representa o mapa do Bing Maps, identifica a mesma rua como "Rua da Agra Nova". Ambos identificam a rua perpendicular a esta como "Rua da Agra Nova". O que ocorre nestes casos é que existe uma diferença no local onde a "Rua Agra Nova" termina para cada um dos mapas, sendo apresentadas por estes diferentes fronteiras entre as duas ruas em questão. Neste caso não temos como definir qual a rua correta; no entanto, pode ser medida a distância entre estas duas ruas e é possível verificar neste caso que a distância entre os centros das duas ruas é apenas 130 metros.

Para medir esta distância deve-se pesquisar as coordenadas geográficas para os dois nomes de ruas distintos (tendo o cuidado de utilizar a mesma fonte, GAPI ou BING). Obtidos os dois pontos de coordenadas geográficas, é possível calcular a distância entre eles, através da seguinte fórmula:

$$\begin{aligned} D =& 6371 \times \arccos\left(\cos\left(\Pi\frac{(90-lng_1)}{180}\right)\cos\left(\Pi\frac{(90-lng_2)}{180}\right)+\right.\\ &\left.+\sin\left(\Pi\frac{(90-lng_1)}{180}\right)\sin\left(\Pi\frac{(90-lng_2)}{180}\right)\cos\left(\Pi\frac{(lat_2-lat_1)}{180}\right)\right), \end{aligned}$$

considerando dois pontos de coordenadas geográficas da forma $P_1 = (lat_1, lng_1)$, $P_2 =$

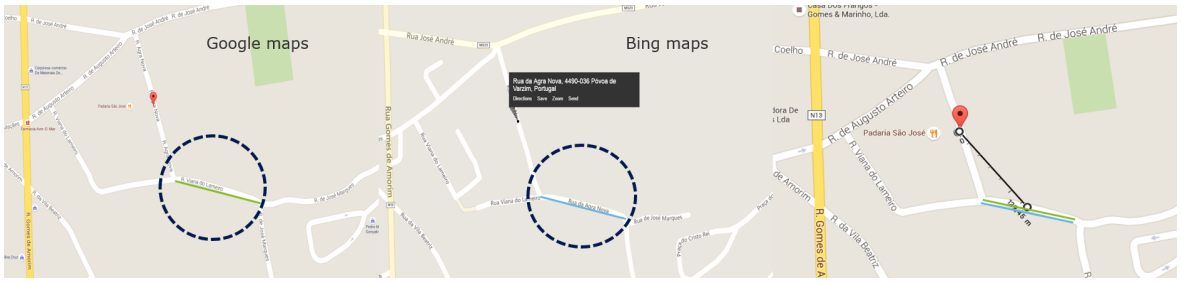


Figura 2.12: Três mapas distintos que apresentam a mesma área. Pode-se observar um mapa do Google Maps com as ruas assinaladas, o mesmo para o Bing Maps e outro que apresenta a distância entre duas ruas.

(lat_2, lng_2). Sempre que esta distância é inferior a 500 metros são considerados os dois nomes corretos e as duas ruas são consideradas hipóteses válidas.

Para a nossa análise este tipo de distância não é problemática, assim sendo podemos considerar que este troço tem os dois nomes associados a si.

Em suma, partindo da base de dados construída pela agregação da informação de três fontes distintas é possível efetuar algumas operações que lhe retiram grande parte dos seus erros. Efetuadas estas operações é necessário combinar a informação restante. Este processo está apresentado na próxima secção.

2.2.3 Comparação de *strings*

Após os tratamentos continuam a verificar-se diferenças entre os nomes das ruas das diferentes fontes. Resta efetuar a comparação destas *strings* com o intuito de detetar erros.

Nesta secção exploram-se os dois métodos de associação criados. O primeiro diz respeito à tentativa de utilização de algoritmos de comparação de *strings* que como se irá verificar não se revelou eficaz neste caso. O segundo tem por base a criação de métodos que visam eliminar diferenças de caracteres que não alteram significativamente o nome.

2.2.3.1 Tratamento 1 - Métodos de comparação de *strings*

O primeiro método consiste em comparar os nomes das ruas, excluindo os espaços, maiúsculas e acentuação. Esta alteração ao texto inicial foi efetuada com o intuito de criar uma maior homogeneização entre as palavras.

Para comparar estes nomes é necessário encontrar um método que quantifique a diferença entre eles. Este método terá de ser efetuado de forma automática, rápida e credível. No Anexo I fez-se a revisão de quatro metodologias de quantificação de diferenças entre nomes. Para desempenhar esta função no algoritmo optou-se pelo nível de Levenshtein. Este valor permite calcular a diferença entre duas *strings* com tamanhos distintos e tem em conta todas as trocas, adições e eliminações de caracteres necessárias para tornar as *strings* iguais. Este método pode ser usado facilmente através de programação pois o Python contém uma biblioteca definida que calcula este valor. Uma operação similar à do nível de Levenshtein é o método de Ratcliff que também contém uma biblioteca em Python que permite a fácil aplicação deste método.

A opção pelo método de Levenshtein prende-se com o facto deste ser mais rápido de calcular pelo Python do que o valor de Ratcliff, o que é uma grande conveniência tendo em conta o elevado número de dados a analisar.

Definiu-se que para relacionar as duas moradas seria utilizada a biblioteca Levenshtein do Python. Esta fornece um valor entre 0 e 1, em que 1 corresponde a textos exatamente iguais. Fazendo a primeira análise de Levenshtein para os nossos dados obtemos um espectro contínuo de valores entre zero e um. Para efeitos práticos é conveniente definir um nível de Levenshtein acima do qual os nomes das ruas serão considerados iguais.

Nível de Levenshtein	Exemplos
1 -> 0.95	Diferença de uma letra. "rua professorleopoldinoloureiro" - "ruaprofessorlepoldinoloureiro"; "ruapadrejoaquimferreira" - "ruapadrejoaquimfereira"
0.95 -> 0.9	Diferença de um ou duas letras, género ou número e ainda casos com "de/a/o/as/os" e "dom". "ruadanielgomesjunqueira" - "ruadomdanielgomesjunqueira"
0.9 -> 0.89	Diferença de palavras ou uma das anteriores. "ruapadrealexandrinoleituga" - "ruaalexandrinoleituga"
0.89 -> 0.85	Nomes diferentes ou palavras extra. "casadospoveiros" - casadospoveirosdorio""; "rampadacabreira" - "rampadacraveira"; "ruadopadrejoaojosedacruz" - "ruadopadrejoaodacruz"
0.85 -> 0.8	Mesmos casos que anteriormente, alterando o tamanho do nome pelo que os erros ganham maior relevância. "ruasacrafamilia" - "ruadasagradafamilia"
0.8 -> 0.75	Igual ao anterior. "rua31dejaneiro" - "travessa31dejaneiro"
0.75 -> 0.7	Igual ao anterior. "travessadetras" - "vieladetras"
<0.7	Ruas bastante diferentes. "ruadaaranha" - "ruadafontainha"; "ruadeentreparedes" - "ruadetras"

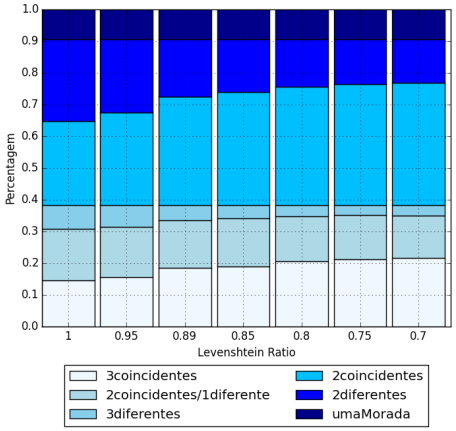
Tabela 2.3: Exemplos de moradas avaliadas para cada gama de valores de Levenshtein.

A tabela 2.3 mostra alguns exemplos de casos avaliados e os respetivos valores de Levenshtein obtidos. Durante este trabalho foram efetuados vários testes através deste método. Inicialmente parecia acertado considerar o nível de Levenshtein igual ou superior a 0.89 como correto. No entanto encontrar este limite revelou-se bastante difícil. Ou incluímos como corretos muitos nomes de ruas claramente diferentes ou excluímos ruas que apresentam o mesmo nome à parte de um erro.

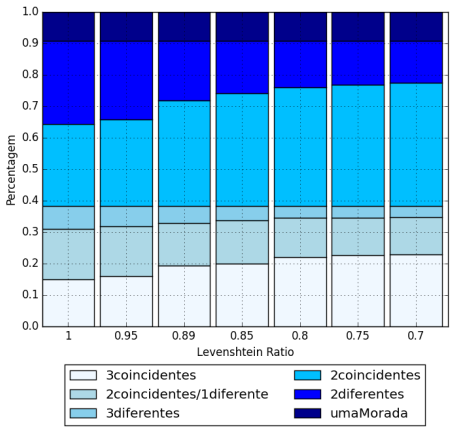
Tendo em conta que este método não se mostra muito eficaz na resolução do problema é necessário tentar classificar melhor os dados. Esta abordagem foi efetuada no tratamento 2, descrito na subsecção seguinte.

2.2.3.2 Tratamento 2 - Eliminação de palavras chave

Este método passa por analisar a informação disponível de forma mais exaustiva e avaliar as alterações que podem ser feitas para eliminar fontes de diferença irrelevantes. Para este efeito



(a) Variação da percentagem de troços que corresponde a cada condição em função do valor de Levenshtein. Na construção deste histograma foram utilizados os nomes de ruas sem espaços.



(b) Variação da percentagem de troços que corresponde a cada condição em função do valor de Levenshtein. Neste caso, para a construção do histograma foram utilizados os nomes de ruas com espaços.

Figura 2.13: Percentagens do nível de Levenshtein para diferentes alterações da morada.

categorizam-se os dados de forma a verificar a percentagem de troços que estão resolvidos.

Considerou-se que existem 6 categorias distintas que são classificadas como:

- x1: três fontes sendo as três moradas iguais;
- x2: tês fontes sendo duas iguais e uma diferente;
- x3: três fontes sendo as três diferentes;
- x4: duas fontes iguais;
- x5: duas fontes diferentes;
- x6: apenas uma fonte.

Dividir os troços em categorias permite verificar a percentagem de troços resolvidos. É importante analisar estas percentagens a cada alteração efetuada pois assim é possível avaliar os resultados obtidos.

Uma dúvida existente prendia-se com o facto de tirar ou não os espaços entre as palavras, tal como efetuado no tratamento 1. Na tentativa de perceber qual a melhor solução, representou-se o histograma com a variação de casos resolvidos para diferentes níveis de Levenshtein. Representou-se um histograma utilizando espaços e outro sem os mesmos. Da figura 2.13a e da figura 2.13b podemos verificar que embora não difiram muito e no caso do valor de Levenshtein ser igual a um serem exatamente iguais, por exemplo para o valor do nível de Levenshtein igual 0.8 o número de troços que pertencem à condição x1 (três fontes iguais) é maior do que no caso em que não têm espaços.

Outro pormenor importante é o facto de muitos erros existentes estarem relacionados com as preposições "de/a/o/os/as". Se pensarmos em extrair estas palavras para diminuir estes erros, temos de ter espaços, caso contrário, sempre que um nome contivesse alguma destas

palavras essa parte da palavra seria eliminada. Assim sendo passa-se a utilizar moradas com espaços. No entanto mantem-se a exclusão de acentos.

Foram testadas uma série de hipóteses fazendo uma análise aos casos que contêm moradas do GAPI e do BING com o intuito de se perceber o que difere entre os nomes que não coincidem. Foi avaliado neste processo a variação do nível de Levenshtein sendo criadas tabelas para verificar como variam os resultados face às diferentes mudanças efetuadas nas moradas.

Alteram-se os nomes segundo um determinado conjunto de parâmetros e verifica-se qual o número de troços que pertence a cada nível de Levenshtein para as diferentes alterações. O conjunto de parâmetros a utilizar é o seguinte:

A: Palavras 'tipo' que se encontram em falta, foram consideradas "dom", "padre", "nossa", "doutor" e "nova";

B: Nomes exatamente iguais no entanto diferem de uma palavra como por exemplo "joão", "dias", neste caso são palavras aleatórias;

C: Erros como por exemplo troca de j/g, plural ou género;

D: Letras trocadas mas que podem significar uma palavra diferente: ferreira/pereira;

E: Palavras trocadas: manuel/joao;

F: Troca do tipo de via: rua/travessa; rua/avenida;

G: Falta de um 'de/da/do/das/dos';

H: Falta de um espaço entre palavras;

I: Estradas nacionais;

J: Palavras escritas em ordem trocada;

K: Ruas diferentes: 'rua da fontainha', 'rua da praia'.

Na tabela 2.4 podemos verificar a análise efetuada a 1667 casos exemplo, extraídos dos dados da Póvoa de Varzim. Na primeira coluna apresentam-se os níveis de Levenshtein, seguidos do número total de casos que corresponde a cada uma das condições. As colunas restantes dizem respeito às condições indicadas acima e mostram quais as diferenças encontradas nas moradas que não são iguais. Através de uma tabela como esta é possível criar uma ideia mais concreta sobre as diferenças presentes nas nossas moradas.

Podemos observar na tabela 2.4 que para casos em que o nível de Levenshtein toma valores entre 0.9 e 1, não inclusive, as principais diferenças prendem-se com erros de preposições, nomeadamente a diferença entre de/da/do/das/dos e ainda erros de escrita: troca entre "g" e "j"; diferenças de género; diferenças de grau. Nestes casos torna-se difícil avaliar qual a versão correta.

Levenshtein	Total	A	B	C	D	E	F	G	H	A+G	B+G	I	J	K
=1	713	-	-	-	-	-	-	-	-	-	-	-	-	-
<1-0.95	41	-	-	20	2	-	-	17	2	-	-	-	-	-
0.90-0.95	141	5	1	38	6	1	-	89	1	-	-	-	-	-
0.85-0.90	45	10	4	4	1	2	7	13	-	1	3	-	-	-
0.80-0.85	46	1	4	-	1	4	17	-	-	4	7	-	-	-
0.75-0.80	23	1	-	-	-	-	19	-	-	-	1	1	1	-
0.70-0.75	21	4	4	-	-	1	2	-	-	1	-	4	1	4
0.65-0.70	13	-	2	-	-	-	2	-	-	-	-	1	-	8
0.60-0.65	44	1	2	-	-	-	-	-	-	-	-	10	-	31
<0.60	580	-	-	-	-	-	-	-	-	-	-	-	-	-

Tabela 2.4: Tabela com os troços catalogados segundo as diferenças entre os mesmos.

Exclui-se de todas as moradas as preposições que se consideram fonte de erro e construiu-se uma nova tabela, idêntica à anterior mas com as moradas alteradas. Pode-se verificar na tabela 2.5, quando comparada com a tabela 2.4 que o número de casos com nível de Levenshtein igual à unidade passa de 713 para 849. O número de casos abaixo do nível 0.6 também aumenta. A ideia passa precisamente por diminuir os valores duvidosos, ou seja entre 0.6 e 0.99 o máximo de moradas possíveis.

Levenshtein	Total	A	B	C	D	E	F	H	I	J	K	A+F	B+E
=1	849	-	-	-	-	-	-	-	-	-	-	-	-
<1-0.95	29	-	-	23	3	-	-	3	-	-	-	-	-
0.90-0.95	31	5	1	17	5	1	-	2	-	-	-	-	-
0.85-0.90	47	19	12	1	2	5	7	-	-	-	1	-	-
0.80-0.85	24	6	5	-	1	1	9	-	-	1	1	-	-
0.75-0.80	15	1	-	-	-	-	13	-	-	1	-	-	-
0.70-0.75	25	3	3	-	-	-	12	-	4	1	-	1	12
0.65-0.70	18	1	4	-	-	-	6	-	1	-	6	-	-
0.60-0.65	40	1	2	-	-	-	1	-	10	-	26	-	-
<0.60	589	-	-	-	-	-	-	-	-	-	-	-	-

Tabela 2.5: Tabela com os troços catalogados segundo as diferenças entre os mesmos aplicando a condição de excluir todos os de/a/o/as/os.

Outro dos erros com que nos deparamos é por exemplo a existência de palavras extra, tal como "dom" ou "padre", tendo casos em que a única diferença é esta palavra.

Determinadas condições de mudança foram estipuladas e aplicadas a cada uma das palavras antes de as comparar. As alterações efetuadas foram as seguintes:

- Excluir preposições (de; do; da; das; dos);
- Excluir a palavra rua apenas se após esta alteração as moradas ficam iguais;
- Excluir as palavras da seguinte lista: "dom", "padre", "nossa", "doutor" e "nova";
- Alterar as palavras de plural para singular nos casos em que esta operação faz com que as moradas sejam iguais;

- Excluir uma a uma as palavras da morada e comparar a cada exclusão, esta operação começando apenas da segunda palavra.

Para observar o efeito das transformações propostas na comparação das moradas, representou-se na figura 2.14 o número de casos com um determinado nível de Levenshtein antes e depois das alterações. Olhando para o canto superior direito, pode-se observar que muitos casos que inicialmente tinham um nível de Levenshtein entre 0.8 e 1, após as transformações, passam para o valor de 1. Existe uma região de valores de Levenshtein iniciais entre 0.4 e 0.7 que após a transformação diminui (formando uma nuvem de pontos abaixo da diagonal). É possível concluir que esta alteração prévia das moradas é vantajosa para o nosso estudo.

O método utilizado impõe que o nível de Levenshtein seja igual a um. Então começamos com duas moradas, aplicamos a ambas todas as condições expostas acima e de seguida comparamos as duas palavras alteradas, obrigando a que sejam exatamente iguais. Caso estas sejam iguais então está encontrado o nome para o troço, caso contrário esse é considerado sem solução. Não esquecer que para o nosso estudo será mais vantajoso ter uma menor quantidade de casos resolvidos do que deixar que o modelo final contenha moradas erradas, pois isso poderá influenciar o resultado final.

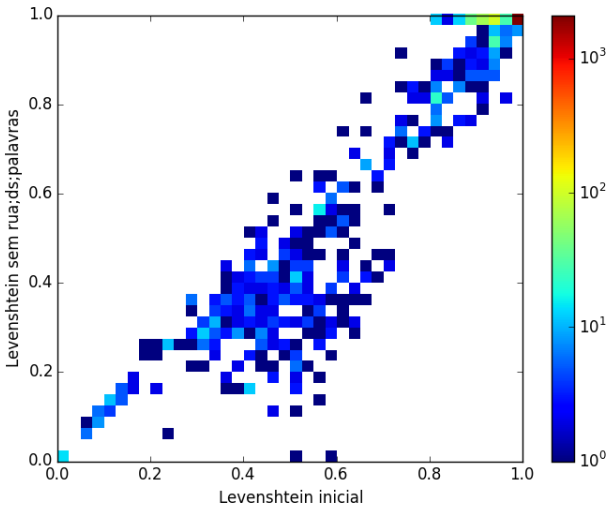


Figura 2.14: Gráfico que representa o nível de Levenshtein.

Os tratamentos 1 e 2 foram criados e otimizados tendo em conta os dados da Póvoa de Varzim de forma a resolver o maior número de casos possíveis e da forma mais correta. Após a aplicação do tratamento 2, o programa foi testado também para a cidade do Porto sendo os resultados obtidos também positivos, nomeadamente 94.6% de casos resolvidos para a Póvoa de Varzim e 92.44% para o Porto.

2.3 CTT

Partindo da solução descrita anteriormente, onde associamos a cada troço as moradas provenientes do OSM, GAPI e BING, pretende-se adicionar a informação relativa ao código postal correspondente. Os nomes de ruas, como vimos pelo tratamento anterior suscitam muitas dúvidas devido aos erros associados. A forma mais exata de definir uma morada é o código postal de sete dígitos. Para associar este código à morada usa-se a lista dos CTT que contém todos os códigos postais do país e a respetiva rua, número de porta e freguesia.

Nome da rua	Campo extra	Número da porta	Código postal	Freguesia
rua da igreja	-	-	4570-033	balazar
rua da igreja	-	-	4570-327	laundos
rua da igreja	-	-	4495-213	navais
rua da igreja	agucadoura	-	4495-027	povoa de varzim
rua da igreja	-	-	4490-517	povoa de varzim
rua da igreja	beiriz	-	4495-332	povoa de varzim

Tabela 2.6: Excerto da tabela dos CTT com ruas iguais para diferentes moradas.

A tabela dos CTT contém a seguinte informação:

- Nome da rua;
- Campo extra: este campo por norma encontra-se vazio, no entanto pode estar associado a freguesias, bairros ou até nomes de ruas;
- Número da porta: quando está preenchido contém intervalos de números;
- Código postal: código de sete dígitos associado à morada;
- Freguesia: freguesia na qual a morada está inserida.

Para associar o código postal à morada é necessário comparar os nomes de ruas presentes na base de dados com o nome de rua dos CTT. Existe uma série de particularidades a ter em conta nesta associação: combinar apenas o nome da rua não é suficiente para criar uma relação unívoca. Nas subsecções seguintes serão explorados estes detalhes.

2.3.1 Campo Extra dos CTT

Um dos problemas da procura na tabela dos CTT é a freguesia. Nem sempre a freguesia que está presente na base de dados corresponde ao nome contido no campo freguesia da tabela dos CTT.

Na tabela 2.6 está presente um excerto da tabela dos CTT. Se for analisada esta tabela verifica-se que existem seis ruas, todas da Póvoa de Varzim, que apresentam o mesmo nome. É necessário distinguir de alguma forma estas ruas, sendo que a diferença entre elas é a freguesia. O problema surge porque três delas aparecem com a mesma freguesia, no entanto o campo extra em duas delas está preenchido. Isto acontece pois a freguesia no caso de aguçadoura e de beiriz está indicada no campo extra.

É necessário definir como se irá efetuar esta associação entre a base de dados com os nomes das ruas e as moradas dos CTT. Se escolhermos como condições que o nome da rua e que a freguesia têm de ser iguais, quando a morada "rua da igreja" com a freguesia "Póvoa de Varzim" efetuar a pesquisa na tabela dos CTT, esta irá encontrar três moradas iguais e assim associar três códigos postais (4495-027; 4490-517; 4495-332). Neste caso não é suficiente definir a freguesia através do campo freguesia dos CTT, sendo necessário definir uma forma de associar o campo da freguesia e o campo extra da tabela 2.6.

Temos de ter em atenção que nem sempre as freguesias obtidas pelo GAPI e pelo BING são iguais à freguesia contida no campo freguesia dos CTT. No caso de Aguçadoura por exemplo, a tabela dos CTT está preenchida no campo freguesia com o nome "Póvoa de Varzim" e em casos nos quais a mesma rua existe na Póvoa de Varzim e em Aguçadoura a tabela contém um campo extra preenchido como "Aguçadoura". O GAPI e o BING por norma para Aguçadoura retornam a freguesia Póvoa de Varzim. No entanto e apenas para o GAPI, a página em formato *json* obtida através do *link* de pesquisa indicado no capítulo 2.2.1 contém informação extra. No GAPI existem três níveis de freguesia em que para o caso de Aguçadoura aparecem como "Póvoa de Varzim", "Aguçadoura" e "Aguçadoura". A primeira freguesia que aparece pode ser o nome da freguesia ou da cidade como é o caso, o segundo nome é o nome da freguesia e o terceiro pode ser o nome da freguesia ou o nome de uma área da freguesia. São recolhidas então as duas primeiras freguesias do GAPI para assegurar que se consegue correspondência nos casos em que os CTT contêm ou não o campo extra preenchido.

É necessário procurar os que contêm a freguesia no campo extra dos CTT em primeiro lugar e só depois procurar os que têm a freguesia no campo freguesia, ou seja o campo extra dos CTT vazio.

2.3.2 Comparar diferentes fontes

Outra informação que pode estar contida no campo extra é o nome de um bairro. Nestes casos a tabela dos CTT apresenta o campo extra preenchido com o nome do bairro no qual o troço se encontra. Para conseguir encontrar o código postal relacionado com esta rua é necessário utilizar no mínimo duas fontes de informação. Como podemos observar na tabela 2.15 que representa um excerto de uma tabela dos CTT, para encontrar um destes códigos postais é necessário relacionar a tabela 2.17 e a tabela 2.16. Quando o campo extra dos CTT contém o nome de um bairro é necessário combinar a informação de duas fontes para chegar a um código postal.

A título de exemplo pensemos nas tabelas 2.15, 2.16 e 2.17 que representam um caso real tratado neste projeto. Existe uma forma de pesquisa que parece numa primeira visualização imediata: procurar o nome da rua fornecido pelo BING na base de dados dos CTT. No entanto existem vários bairros com o mesmo nome de rua e pertencentes à mesma freguesia. É necessário combinar a informação do GAPI para se perceber qual o bairro em questão. A pesquisa nestes casos terá de ser diferente do que foi visto na subsecção 2.4.1 em que seria procurada a informação pela sua morada. Neste caso, será necessário procurar com duas moradas em que uma delas será o nome do bairro e outra o nome da rua.

Estes casos correspondem aos resolvidos em 2.2.2.4 pois por norma quando é efetuada a procura do nome "rua 1, 4490-168" e "bairro nova sintra, 4490-552" estes encontram-se a uma distância inferior a 500 metros. Segundo o critério atrás apresentado são consideradas as duas moradas válidas. Como são guardadas as duas moradas é possível depois efetuar a pesquisa nos CTT com estas duas fontes.

Nome da rua	Campo extra	Nr da porta	CodPost	Freguesia
rua de antonio batista de almeida	bairro nova sintra	-	4490-468	povoa de varzim
rua 1	bairro nova sintra	-	4490-168	povoa de varzim
rua 2	bairro nova sintra	-	4490-170	povoa de varzim
rua 3	bairro nova sintra	-	4490-172	povoa de varzim
rua 4	bairro nova sintra	-	4490-696	povoa de varzim

Figura 2.15: Tabela ilustrativa de um caso de ruas pertencentes a um bairro. Excerto da tabela dos CTT

ID	osm_id	Rua_GAPI	Nr da Porta (GAPI)	CodPost e Freg. (GAPI)	Freguesia (GAPI)
184	32726843	bairro de nova sintra	22-23	4490-552 povoa de varzim	povoa de varzim
185	32726844	bairro de nova sintra	85-86	4490-483 povoa de varzim	povoa de varzim
188	32726848	bairro de nova sintra	38	4490-552 povoa de varzim	povoa de varzim
190	32726848	bairro de nova sintra	1	4490-552 povoa de varzim	povoa de varzim
191	32726855	bairro de nova sintra	26-27	4490-552 povoa de varzim	povoa de varzim

Figura 2.16: Excerto da tabela do GAPI.

ID	osm_id	Rua_BING	Nr da Porta (BING)	CodPost e Freg. (BING)	Freguesia (BING)
184	32726843	rua 2	61	4490-170 povoa de varzim	povoa de varzim
185	32726844	rua 3	97	4490-172 povoa de varzim	povoa de varzim
188	32726848	rua 4	39	4490-490 povoa de varzim	povoa de varzim
190	32726848	rua antonio baptista almeida	11	4490-468 povoa de varzim	povoa de varzim
191	32726855	rua 1	26	4490-168 povoa de varzim	povoa de varzim

Figura 2.17: Excerto da tabela do BING.

2.3.3 Número da porta

Em ruas com grande comprimento é frequente o código postal mudar com o número da porta. Este comportamento constitui uma dificuldade adicional na automatização da pesquisa. Nestes casos, o campo da tabela dos CTT com o número da porta encontra-se preenchido. É necessário que o método utilizado verifique que esta entrada está presente. Nesse caso é necessário verificar que o número de porta da fonte se encontra na lista de números possíveis dos CTT (os números são ordenados e por norma estão separados entre pares e ímpares).

2.3.4 Programa desenvolvido

Para combinar a informação necessária, tendo em conta todas as restrições existentes, o método encontrado passa por estabelecer a prioridade das operações, sendo estas:

1. Quando o campo extra dos CTT indica um nome de rua ou o nome de um bairro temos de combinar a informação do nome da rua e do campo extra dos CTT com os nomes das duas fontes;

2. Quando o campo extra contém informação sobre a freguesia esta tem de ser combinada com a freguesia da fonte;

3. Se o campo extra estiver vazio fazemos a comparação habitual dos nomes de ruas.

- De notar que para cada grau de prioridade é necessário verificar se a morada dos CTT contém número de porta.

O código é elaborado de forma a que exista uma variável que define a prioridade da comparação. Começa-se pela primeira morada da fonte, que imaginemos que é "rua da igreja, aguçadoura, póvoa de varzim" então essa morada é comparada com cada uma das moradas da lista dos CTT; encontrada a morada "rua da igreja, póvoa de varzim" esta irá ser guardada e define-se a prioridade com o nível 3; o programa irá continuar a percorrer a lista dos CTT e de seguida encontra a morada "rua da igreja, aguçadoura, póvoa de varzim" então a morada anteriormente guardada é substituída pela morada agora encontrada e o nível de prioridade passa para 1; caso volte a encontrar uma morada em que apenas o nome e a freguesia sejam iguais, como o nível de prioridade é 3, ou seja, maior do que o atual então essa morada é ignorada.

Depois de todo este tratamento verificou-se que cerca de 30% dos códigos postais não eram atribuídos a nenhum troço. Isto acontecia devido à falta de estradas do OSM e a resolução é explorada na próxima subsecção.

2.4 Encontrar Moradas

O último passo foi efetuado na tentativa de apurar o motivo pelo qual cerca de 30% dos códigos postais não foram atribuídos a nenhuma rua. Após a análise de alguns casos verifica-se que nem todas as ruas apresentam um troço associado no OSM. Na figura 2.18 existe uma rua contida dentro do círculo que não é representada pelo OSM (ruas a azul) e como tal o seu ponto central não consta da lista inicial de troços que tratamos ao longo do capítulo 2.

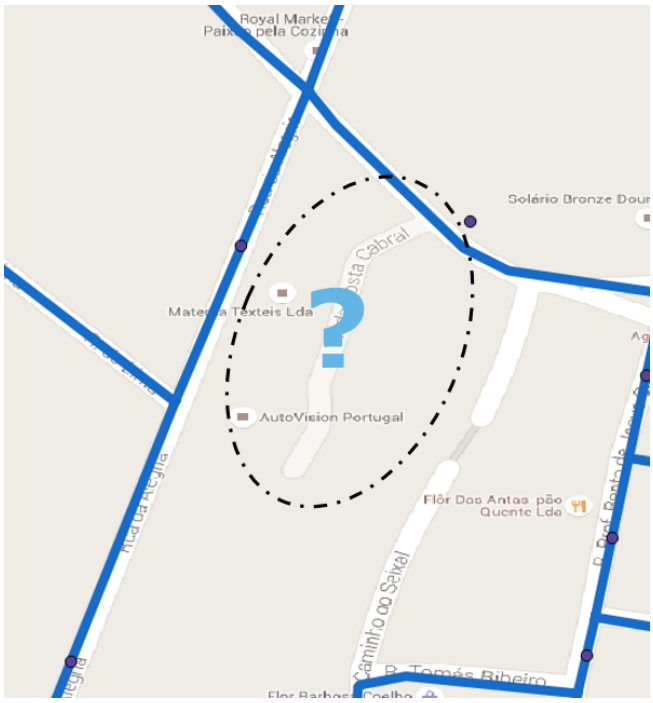


Figura 2.18: A imagem de fundo faz parte do Google Maps e as linhas azuis representam os troços do OSM.

A solução para o problema em questão passa por tentar atribuir o código postal da rua que não está representada à rua mais próxima. Para resolver esta questão utiliza-se a lista dos CTT e recolhem-se todas as moradas cujo código postal não está atribuído a nenhum troço. A partir desta lista, e através do GAPI e depois do BING, em passos separados, procuram-se as coordenadas geográficas associadas a cada uma das moradas.

As informações inicialmente disponíveis são as coordenadas geográficas de cada um dos troços do OSM (a partir das quais obtivemos a morada de cada troço). Para cada elemento dos CTT que não se encontra associado a nenhuma morada procura-se a coordenada geográfica dos troços iniciais mais próxima (através do cálculo da distância efetuado da mesma forma que na subsecção 2.2.2.4). No caso da morada mais próxima se encontrar a menos de 500 metros de distância é associado o código postal ao troço.

Esta operação é aceitável neste estudo tendo em conta que o objetivo é a caracterização de cidades a nível geográfico e não de ruas em particular (nesse caso este passo não era possível).

Este tratamento foi desenvolvido e otimizado tendo em conta as estradas da Póvoa de Varzim e chegou-se a um resultado final com um total de 94.31% dos casos resolvidos.

Após alcançados resultados satisfatórios para o local de teste, verificou-se o resultado para o Porto com o intuito de se perceber se este código poderia ser utilizado para tratar todo o país. Para a cidade do Porto atingiu-se 92.44% dos troços com morada e código postal associados.

No mapa da figura 2.19 podemos observar os resultados obtidos pelo algoritmo para a cidade do Porto. A verde estão representados os troços encontrados e a azul aqueles para os quais não foi possível associar um código postal.

Com este mapa é possível efetuar uma análise estatística dos dados e uma análise geográfica. Deste modo associa-se um cliente a uma posição no mapa, de forma rápida e eficaz, processo que de outra forma não poderia ser efetuado.

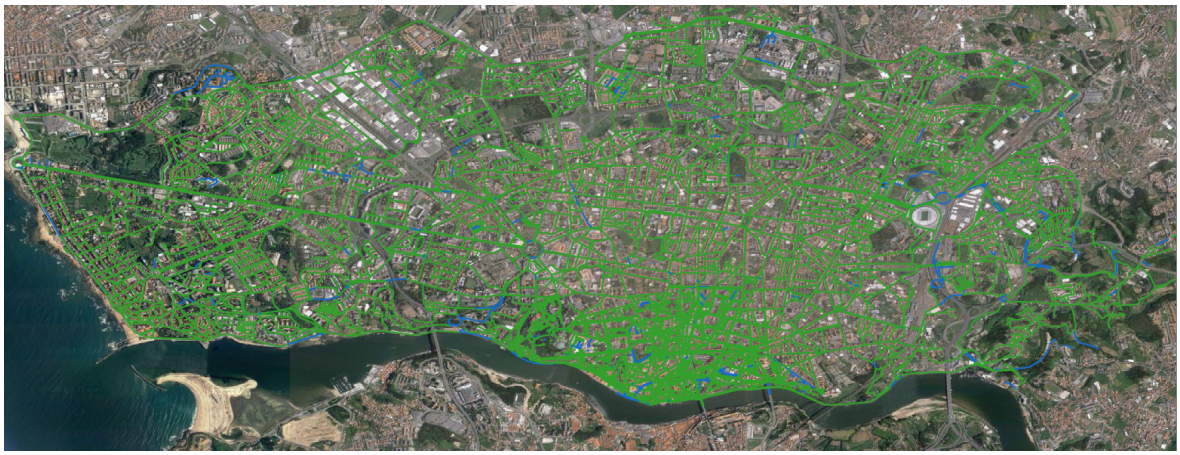


Figura 2.19: Mapa final do Porto. Encontram-se assinalados a verde os troços resolvidos e a azul aqueles para os quais não se obteve nenhum código postal.

Os dados fornecidos pela Deloitte sobre clientes contêm cerca de 75% das moradas com um código postal válido. Dos clientes que contêm um código postal válido 97% são associados a uma morada. Para os 25% dos casos em que não existe código postal válido, apenas 35% dos clientes são possíveis de associar a uma morada no mapa.

3 Análise dos Resultados

Para este estágio foi disponibilizada informação sobre uma dada entidade bancária por parte da empresa. Na figura 3.1 apresentamos um resumo com a descrição dos dados iniciais fornecidos. Nos dados existem informações pormenorizadas por cliente com a morada, idade, habilitações literárias e valores de crédito e depósito associados à instituição bancária. Para os tratamentos efetuados neste capítulo são usados os códigos das habilitações presentes na figura 3.1. A partir destas informações e de dados de fontes livres foi criado um SIG que possibilita o estudo geográfico de informação partindo de moradas, tal como descrito no capítulo 2.

Partindo destas duas fontes é possível efetuar dois tipos de análise, estatística e geográfica. Pode ser efetuada uma análise estatística dos resultados apenas com os dados da entidade bancária ou uma análise geográfica se utilizarmos também o SIG criado neste projeto.

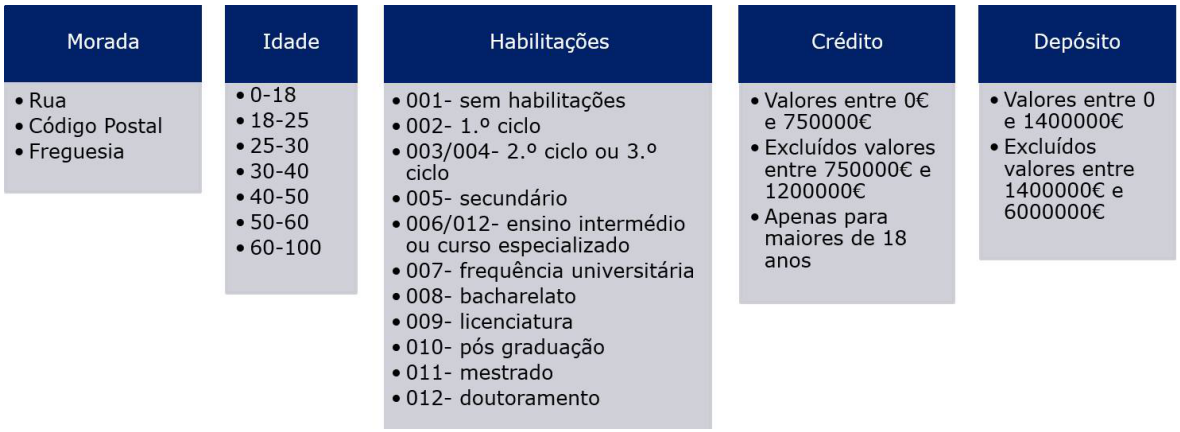
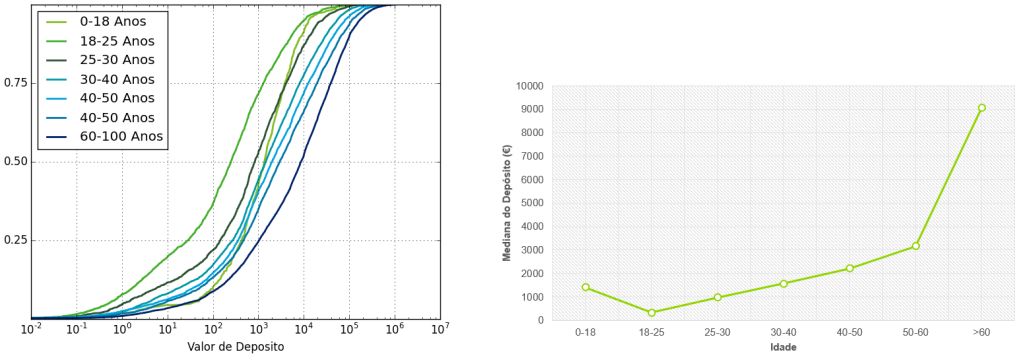


Figura 3.1: Informação da entidade bancária fornecida pela empresa. Pode-se verificar as diferentes classificações permitidas para as habilitações bem como os valores de créditos e depósito presentes nos nossos dados.

Nas duas secções seguintes iremos efetuar estas análises para os dados disponíveis, restringindo o estudo à cidade do Porto.

3.1 Análise Estatística

Esta secção contém uma análise estatística da informação dos dados descritos na figura 3.1. Tendo em conta os dados disponíveis e a natureza de uma entidade bancária é de interesse verificar a variação dos valores de depósito e crédito com dados como idade e habilitações.



(a) Curvas cumulativas do valor de depósito para diferentes faixas etárias.

(b) Mediana do depósito em função da idade dos clientes.

Figura 3.2: Valores de depósito em função da idade.

3.1.1 Valor do depósito em função da idade

É de esperar que o valor de depósito aumente com a idade. Para analisar esta dependência construíram-se as curvas de probabilidade cumulativa do valor de depósito para diferentes faixas etárias.

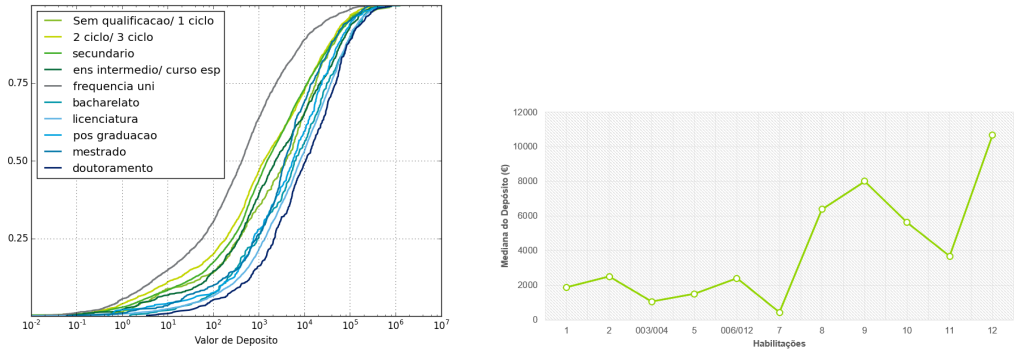
Através da figura 3.2a é possível verificar que todas as curvas seguem a mesma evolução, à exceção da curva para uma faixa etária entre os 0 e os 18 anos. Para esta faixa etária existe uma subida muito abrupta, estando os valores concentrados em dois extremos. Este facto pode estar relacionado por um lado com pais que abrem a conta para os filhos e colocam um valor pontual (por exemplo dinheiro do batizado) e não alteram mais o valor da conta ficando sempre abaixo dos 100€, ou por outro lado por exemplo avós que preferem passar heranças diretamente para os netos e tendo assim valores superiores a 5000€.

Para as restantes faixas etárias verifica-se que as curvas são deslocadas para a direita com o aumento da idade. O valor de depósito aumenta com a idade, sendo os aumentos mais notórios para a passagem da faixa etária entre os 18-25 para os 25-30 anos e a passagem entre os 40-50 e os 60-100 anos. A faixa etária com menor valor de depósito corresponde a idades entre os 18 e 25 anos, sendo parte desta população estudantes universitários ou pessoas que se encontram em início de carreira.

Se for efetuado um corte no gráfico para o valor de probabilidade de 0.5 e forem extraídos esses valores obtém-se um gráfico da mediana do depósito face à idade, representado na figura 3.2b. Pode verificar-se os resultados das curvas cumulativas evidenciados pelo valor da mediana. Mesmo para idades entre 0 e 18 anos a mediana de depósito é maior do que para intervalo entre 18 e 25 anos. É possível verificar ainda a evidência de um aumento muito acentuado entre os 50-60 anos quando comparado com os maiores de 60 anos, esta condição pode estar associada ao facto de haver uma tendência para que pessoas com maiores rendimentos e habilitações apresentarem maior esperança média de vida [16].

3.1.2 Valor do depósito em função das habilitações

No que diz respeito à variação do depósito face às habilitações literárias a diferença é muito menos notória, tal como se verifica pela figura 3.3a.



(a) Curvas de probabilidade cumulativa do depósito para diferentes habilitações.

(b) Mediana do depósito em função das habilitações.

Figura 3.3: Variação do valor de depósito em função das habilitações.

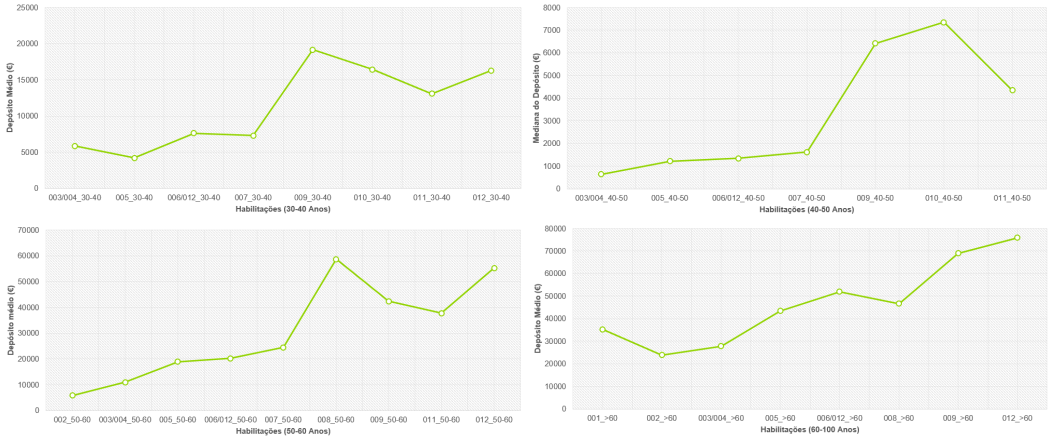


Figura 3.4: Gráficos do valor médio do depósito em função de diferentes habilitações literárias, para diferentes faixas etárias.

Podemos observar que a curva que indica menores depósitos corresponde à curva de frequência universitária o que está de acordo com os dados anteriores, pois estes correspondem maioritariamente à faixa etária entre os 18 e 25 anos.

Pela figura 3.3b parecem verificar-se dois patamares distintos, um para habilitações abaixo do nível 7 (ensino não universitário) e outro acima deste (ensino universitário) que apresentam valores de mediana de depósito maiores.

É importante reter dos gráficos de 3.3 que a variação do depósito não apresenta uma variação bem definida com as habilitações. Para os níveis de habilitações de pós graduação (010) e de mestrado (011), que são níveis mais elevados do que a licenciatura (009), verificam-se valores da mediana de depósito mais baixos. No entanto seria espectável que um maior número de pessoas com mestrado tivesse maior valor de depósito do que pessoas com licenciatura.

Os níveis de habilitações são muito vincados com as faixas etárias em questão. Temos o exemplo de anos posteriores ao tratado de Bolonha em que muitos dos cursos passaram a conter mestrado integrado, aumentando assim o número de pessoas com mestrado. Tendo em conta este facto percebe-se que o valor de depósito em função das habilitações literárias deve ser condicionado para diferentes faixas etárias como pode ser observado nos diferentes

gráficos da figura 3.4.

Pode-se verificar a partir desta figura que entre os 30 e 40 anos temos o valor de depósito mais elevado para os clientes com o grau de licenciatura. Nesta faixa etária a mediana do depósito apresenta valores máximos de 20000€. É compreensível que pessoas nesta faixa etária e com maior grau académico, ou seja, menos anos de trabalho apresentem menores valores de depósito.

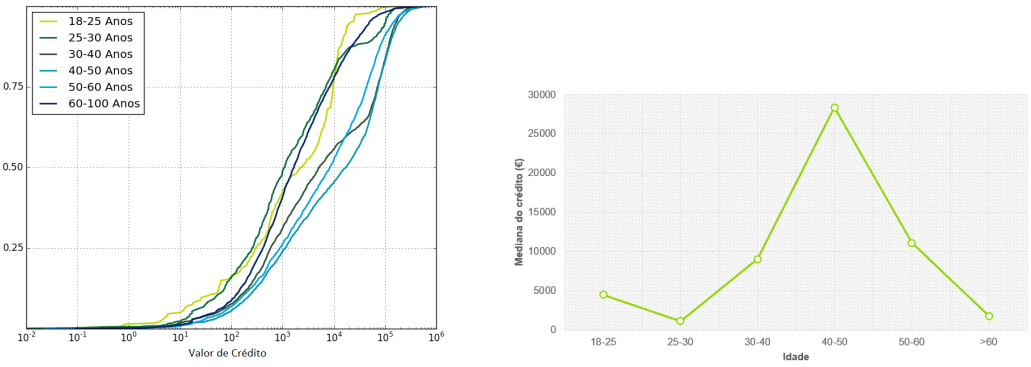
Na faixa etária entre os 40 e 50 anos a forma do gráfico mantem-se aproximadamente igual à faixa etária entre os 30 e 40 anos. Contudo, ao contrário da faixa etária anterior os clientes com grau de mestrado apresentam valores de depósito superior aos que contêm grau de licenciatura. Não se representamos dados sobre pessoas com doutoramento pois este número é muito reduzido.

Na faixa etária entre os 50 e 60 anos verifica-se que o maior valor médio de depósito corresponde ao bacharelato (008) e ao doutoramento (012).

Acima dos 60 anos esta curva tem um crescimento mais homogéneo. Não são apresentados dados de pós graduação (010) e mestrado (011) porque o número de pessoas é muito reduzido.

3.1.3 Valor do crédito em função da idade

Quanto ao crédito é possível, partindo das curvas cumulativas da figura 3.5a, observar que existem zonas de declive mais baixo, ladeadas por zonas de declive mais alto. As regiões de declive acentuado correspondem a valores típicos de crédito e as regiões de declive baixo correspondem a valores raros de crédito.



(a) Curvas de probabilidade cumulativa do valor do crédito para diferentes faixas etárias.

(b) Valor da mediana do crédito em função da faixa etária.

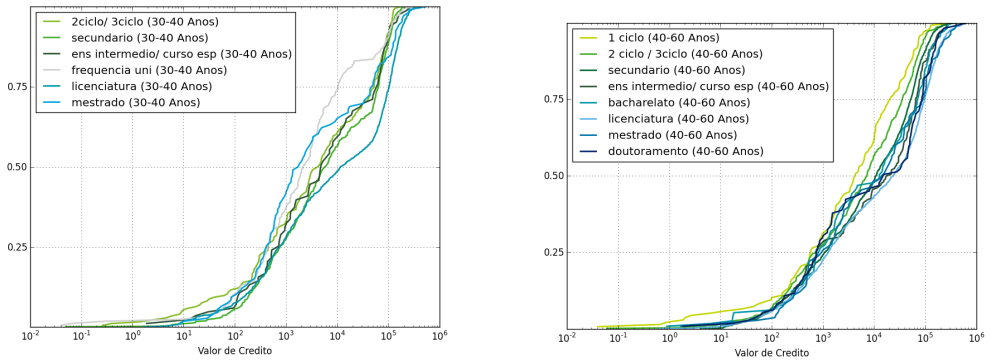
Figura 3.5: Variação do crédito em função da idade.

Analisando a faixa etária entre os 30 e 40 anos é possível verificar que o crédito para a primeira zona maior declive corresponde a valores até ao 10000€ e representa cerca de 55% dos clientes. Estes valores de crédito mais baixo podem ser interpretados como créditos para a compra de produtos de menor valor como eletrodomésticos, viagens ou até a compra de automóveis. A segunda zona de maior declive que corresponde a cerca de 35% dos clientes corresponde a créditos superiores a 70000€ que possivelmente estão associados a créditos à habitação.

O valor da mediana por segmento etário está representado na figura 3.5b. Observa-se que o valor de crédito contraído aumenta até uma gama de idades entre os 40 e 50 anos e de seguida este valor diminui. Este comportamento é o esperado devido à compra de casa. A partir desta faixa etária o valor de crédito diminui com o pagamento mensal das prestações.

3.1.4 Valor do crédito em função das habilitações

A dependência do valor de crédito com as habilitações, não é tão evidentes. Na figura 3.6 é possível observar as curvas cumulativas de crédito em função das habilitações para duas faixas etárias distintas. Partindo destes gráficos verifica-se que quanto maior o nível de habilitações, mais elevada a faixa etária das pessoas que contraem um crédito. Verifica-se que por exemplo para uma faixa etária entre os 30 e os 40 anos, cerca de 50% das pessoas com licenciatura (009), pedem crédito à habitação, enquanto com mestrado (011) apenas 30% das pessoas o fazem.



(a) Gráfico das curvas cumulativas do valor de crédito para cada habilitação para clientes com idade entre os 30 e 40 anos.

(b) Gráfico das curvas cumulativas do valor de crédito para cada habilitação para clientes com idade entre os 40 e 60 anos.

Figura 3.6: Valor de crédito para diferentes níveis de habilitações.

Para a faixa etária entre os 40 e 60 anos, apenas as curvas de mestrado (011) e doutoramento (012) apresentam o patamar mencionado. Este facto não implica que os restantes clientes não tenham créditos, mas que provavelmente já os têm há mais tempo. O efeito da inflação e da amortização faz com que o valor de um crédito mais antigo não tenha um valor típico.

É de notar que embora os valores de crédito sejam aproximados, essencialmente para níveis universitários, parece existir um valor mais elevado para os clientes com licenciatura quando comparados com os restantes.

3.2 Análise Geográfica

O *software* QGIS é uma ferramenta que permite o cálculo de valores médios, variância, desvio padrão, máximos ou mínimos da informação e a sua representação em mapas. É possível condicionar o cálculo destas variáveis às áreas que se pretende tratar. Nós trabalhamos com áreas geográficas definidas no censos pelo que podemos tratar desde regiões a subsecções de freguesias.

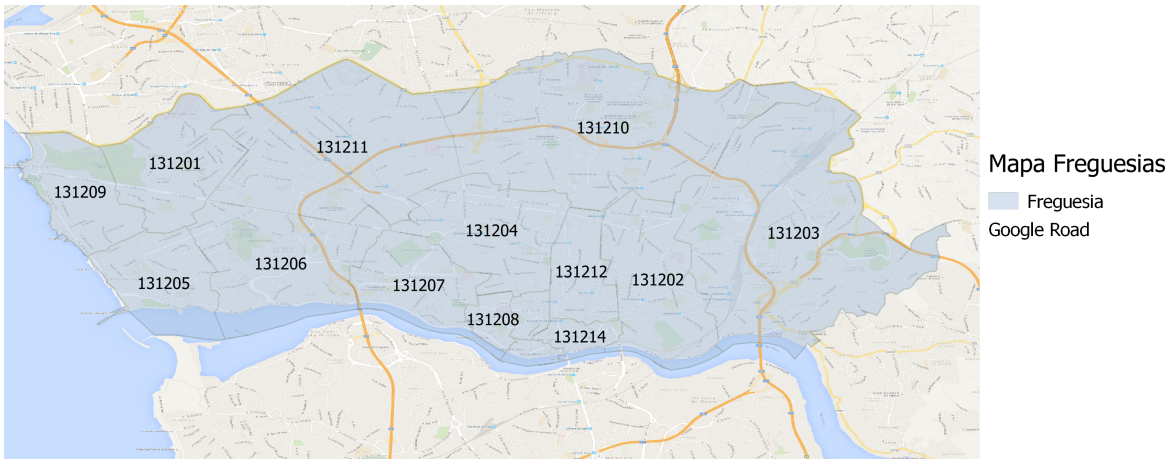


Figura 3.7: Mapa da cidade do Porto, obtido através do Google Maps e mapa das freguesias do recolhido no *site* do INE.

Fazendo uso destas propriedades é possível calcular valores de crédito, depósito ou idades dos clientes para diferentes áreas e tirar ilações dos resultados ou compará-los com outras fontes tal como os dados do INE.

3.2.1 Verificar se a amostra é representativa da população

Tendo em conta que estão disponíveis quer os dados fornecidos pela empresa assim como os dados do INE, é possível avaliar se a amostra é representativa da população, ou seja, se a distribuição de uma dada variável é igual para as duas fontes. Na figura 3.7 estão identificadas as freguesias da cidade do Porto, às quais corresponde um número, sendo os dois últimos dígitos deste os utilizados para definir a freguesia representada nos histogramas que são analisados nesta secção.

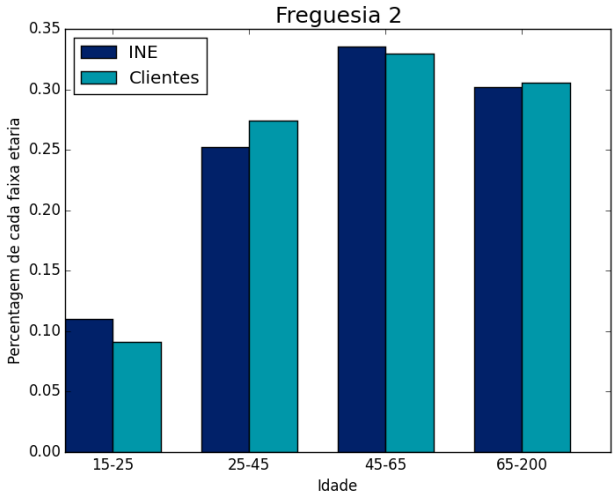


Figura 3.8: Histograma representativo da percentagem de população para diferentes faixas etárias, segundo os dados do banco e os dados do INE.

Representa-se a percentagem de população, de cada uma das fontes, para diferentes faixas etárias em cada uma das freguesias. É possível verificar através do histograma da figura 3.8

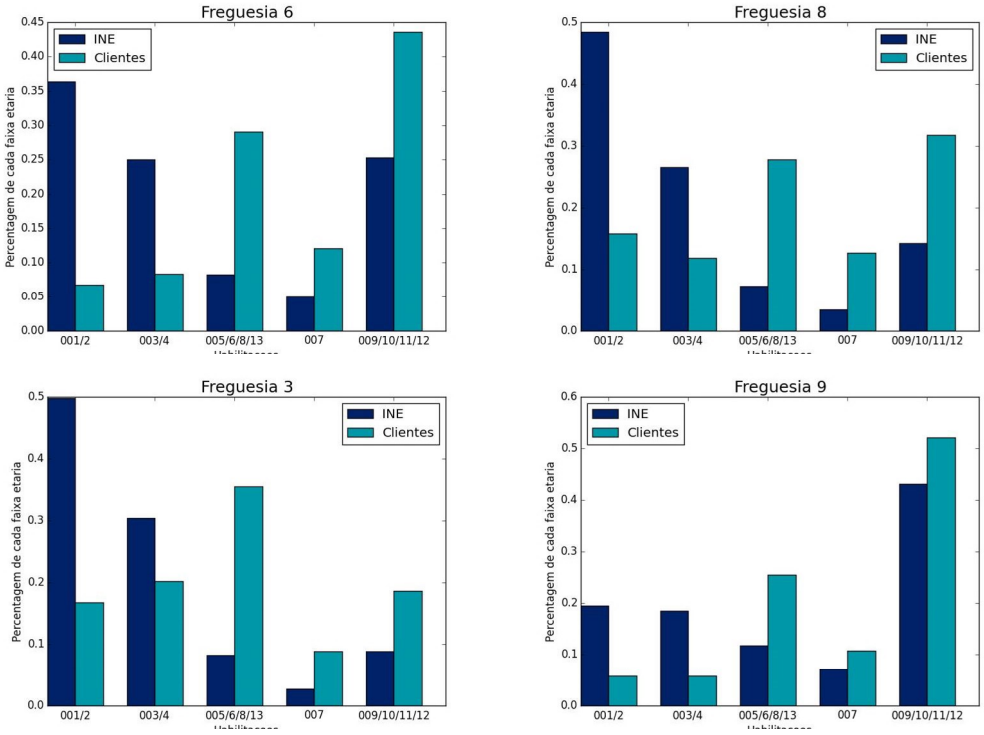


Figura 3.9: Histogramas representativos da percentagem de clientes com cada tipo de habilitação a azul claro e os dados do INE a azul escuro.

que estes dados estão em concordância, ou seja a percentagem de clientes do banco para cada faixa etária é aproximadamente igual às percentagens de população do INE.

Assim pode concluir-se que o banco está a abranger toda a população de igual forma. Através desta análise é possível auxiliar uma entidade bancária a alterar políticas ou verificar se as políticas presentes estão a evoluir no sentido esperado de modo a abranger toda a população por si pretendida. A título de exemplo imaginemos que a percentagem de clientes entre os 25 e 45 anos é muito menor que a percentagem de população segundo o INE. Com esta informação o banco poderá definir uma estratégia de modo a crescer cota nesta faixa etária específica.

A mesma análise pode ser efetuada em relação às diferentes habilitações. Na figura 3.9 podemos observar três amostras obtidas na cidade do Porto. Se for observado o histograma da freguesia 6 verifica-se que segundo o INE esta apresenta cerca de 65% da população com um nível de escolaridade não universitário; no entanto, se fosse efetuada uma análise apenas dos clientes diríamos que apenas cerca de 45% das pessoas não apresenta nível universitário. Os três gráficos da figura representam zonas distintas da cidade do Porto que, claramente e como vemos pelos dados do INE, apresentam diferentes percentagens em cada grau de escolaridade, muito devido a fatores sociais, culturais e económicos. No entanto mesmo na freguesia 9 que apresenta segundo o INE a maior percentagem de população com nível universitário a percentagem de clientes com este nível de habilitações é superior à percentagem do INE.

Destes resultados, conclui-se que a entidade bancária não capta uma grande percentagem de população com baixas habilitações, podendo este facto estar relacionado com uma política

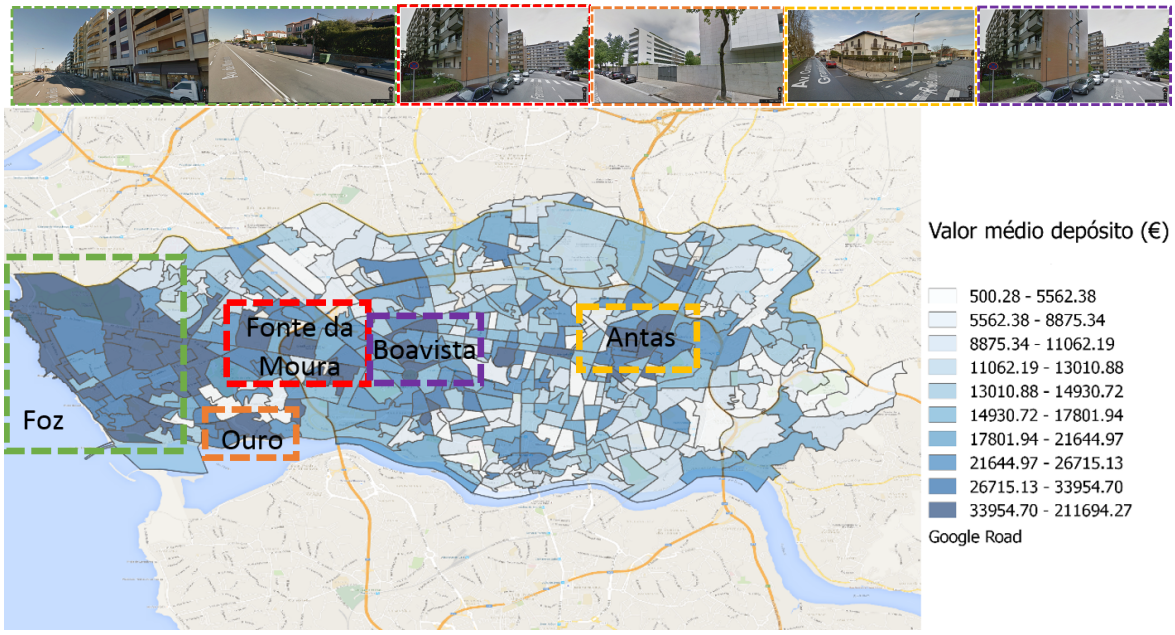


Figura 3.10: No mapa encontra-se representado o valor médio do depósito para cada uma das secções do Porto. Na imagem constam ainda fotografias das principais zonas representadas na figura.

interna do banco, ao qual podem não interessar esses clientes, ou por outro lado, pode estar ligado a uma má captação dos mesmos.

3.2.2 Representação de médias

Uma das possibilidades que o QGIS fornece é o cálculo do valor médio de uma variável escolhida por nós para os pontos que se encontram dentro de cada um dos polígonos, considerando que são representados por pontos os clientes e por polígonos as diferentes secções do Porto.

Na figura 3.10 está representado o valor médio de depósito para cada secção da cidade do Porto. Verifica-se que as zonas com maior valor de depósito coincidem com locais esperados tais como a Foz e as Antas. No mapa podemos observar que entre a zona da Foz e do Ouro existem uns polígonos mais claros, sendo estes representativos de habitações sociais e por isso é de esperar que apresente uma cor mais clara.

Todas as variáveis presentes na informação da *shapefile* podem ser representadas desta forma. No caso dos dados do INE só temos os valores absolutos dentro de cada granularidade. Nos dados fornecidos pela entidade bancária temos um conjunto de clientes em cada polígono, por este motivo para além do valor médio de cada um dos campos podemos calcular e representar correlações, dispersões e toda a gama de medidas estatísticas.

Na próxima subsecção iremos apresentar a medida de uma correlação e a sua representação no mapa.

3.2.3 Correlação

Pode ser avaliada a dependência de duas variáveis. Por exemplo, pode-se verificar se em todas as secções os valores de depósito têm igual dependência com a idade.

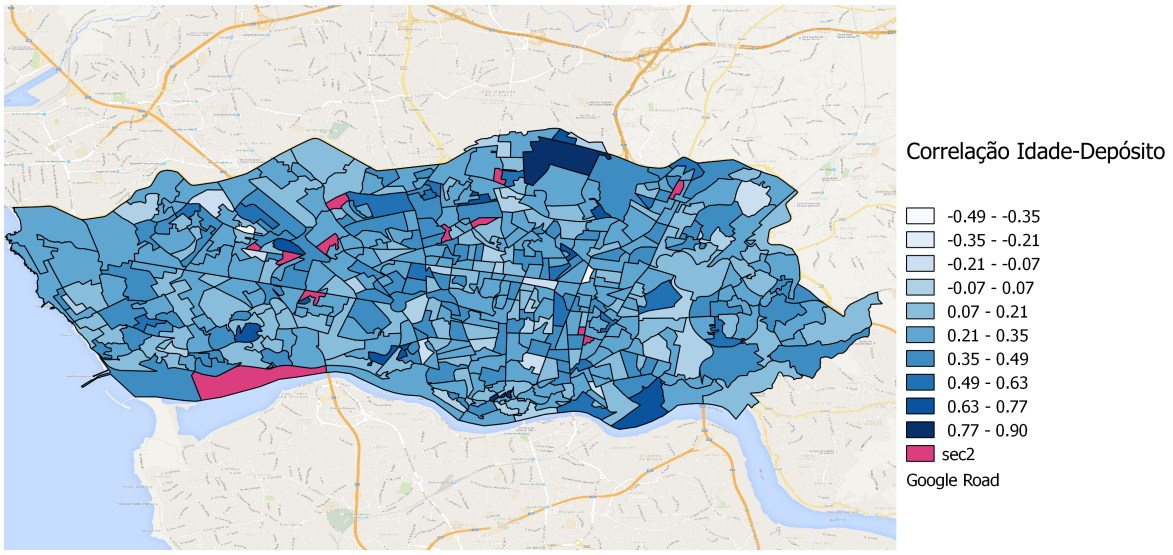


Figura 3.12: Valor do coeficente de correlação entre a idade e o valor de depósito de cada cliente para cada uma das secções.

Uma das formas de avaliar esta relação é através do coeficiente de correlação, nomeadamente o coeficiente de correlação de Pearson, que gera um valor entre -1 e 1, fornecendo assim a dependência de duas variáveis. Quando o este coeficiente toma o valor de -1 indica uma dependência negativa, no caso deste tomar o valor 1 uma dependência positiva e quando é nulo indica a ausência de relação entre as variáveis.

O coeficiente á calculado pela seguinte fórmula

$$\rho_{X,Y} = \frac{\overline{XY} - \overline{XY}}{\sqrt{\overline{X^2} - \overline{X}^2}\sqrt{\overline{Y^2} - \overline{Y}^2}},$$

sendo X e Y as duas variáveis a avaliar.

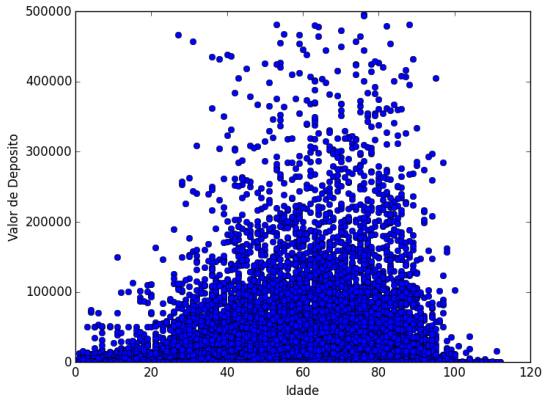


Figura 3.11: Valores de depósito em função da idade.

Caso existam diferenças entre os valores de correlação das diferentes secções é necessário perceber a origem dessas diferenças, as consequências que estas implicam para a entidade e como resolver eventuais problemas que possam estar na base desses resultados.

Na figura 3.12 está representado para cada secção da cidade do Porto o valor do coeficente de correlação entre a idade e o depósito dos clientes.

A correlação para estas duas variáveis varia entre -0.49 e 0.63 para praticamente todas as secções. Segundo os critérios de análise do coeficiente de Pearson estes valores de correlação são considerados baixos.

O valor do coeficiente de correlação de Pearson apresenta um valor bem definido quando

verificado um crescimento dos dados sem grande dispersão. Quando a dispersão dos pontos é muito grande não é possível medir a correlação entre duas variáveis. Para o caso em estudo representamos na figura 3.11 as variáveis idade e valor de depósito. Pode-se observar que a dispersão de valores é muito elevada. Mesmo conhecendo que existe uma tendência para um valor crescente de depósito em função da idade, este fato não se verifica para uma parte dos clientes.

Uma das hipóteses seria limitar a gama de idades por exemplo até aos 45, idade até à qual o valor médio sobe, no entanto a dispersão iria continuar elevada e para além disso quanto mais curto é o intervalo mais difícil se torna medir a correlação devido à diminuição da população estatística.

Não é possível avaliar a dependência crescente entre o valor de depósito e a idade através do cálculo do coeficiente de correlação.

4 Conclusão

Com o desenvolvimento deste projeto verifica-se que é possível criar um sistema de localização geográfica capaz de realizar estudos geográficos a partir de moradas de clientes. É possível construir um processo sistematizado e automático que recorre a fontes de informação livres para atribuir a um mapa nomes de ruas e os respetivos códigos postais. Este processo tem por base um conjunto de processos de recolha e tratamento de informação, nomeadamente de texto, que foram estudados e otimizados para cumprir com o maior sucesso possível a tarefa pretendida. A construção deste modelo torna possível determinar de forma relativamente rápida e precisa, a localização de um cliente.

Sem o processo desenvolvido nesta tese apenas seria possível determinar a localização de um cliente através de coordenadas geográficas. Esta é uma informação pouco usual de obter e de difícil acesso. Com este processo tornou-se possível associar a informação disponível (moradas) às coordenadas geográficas e a sua inclusão num SIG.

A título de exemplo foram tratados alguns dados fornecidos pelo banco ao nível estatístico e ao nível geográfico. O desenvolvimento de um SIG vem complementar o trabalho de análise estatística. É assim possível representar os valores associados a um mapa ou cruzar informações de diferentes mapas, através de operações estatísticas.

Este estágio revelou-se bastante desafiante permitindo compreender a construção de um SIG. Permitiu a aplicação de conhecimentos de *software* lecionados no curso e o contacto com novas ferramentas tal como o QGIS.

O projeto criado apresenta aplicações reais para o mundo empresarial. Neste momento existe uma equipa na Deloitte da qual faço parte para dar continuidade ao trabalho deste estágio.

Bibliografia

[1] Davis JR., C.A. & Fonseca, F.T. Assessing the Certainty of Locations Produced by an address Geocoding System. 2007.

[2] Medeiros, Anderson. (Citação: 20 de Outubro de 2015).Como Desenvolver um GIS – Parte 1.
http://andersonmedeiros.com/como-desenvolver-um-gis1/.

[3] (4 de Novembro de 2015). SIG.
http://www.oern.pt/ver.php?cod=0C0C0B0E.

[4] Chrisman, Nicholas R. What Does 'GIS' Mean? University of Washington : Department of Geography , 1999.

[5] (27 de Outubro de 2015). Manual QGis.
http://docs.qgis.org/2.8/en/docs/user_manual/.

[6] Kurt Menke, Richard Smith Jr, Luigi Pireli, John Van Hoesen. Mastering QGIS. s.l. : Packt Publishing Ltd., 2015.

[7] Goldstein, Stephane. Criação de plataforma de geocoding baseada em serviços Google Maps. Universidade de Lisboa : s.n., 2014.

[8] Silva, Domingos F. P. da. Sistema de informação Geográfica para transportes- uma aplicação aos transportes urbanos de Guimarães. Instituto Superior de Estatística e Gestão da Universidade Nova de Lisboa : s.n., 2006.

[9] Filho, Jugurta L. Introdução A SIG - Sistema de Informação Geográfica. Universidade Federal de Rio Grande do Sul : s.n., 1995.

[10] Instituto Nacional de Estatística. (20 de Outubro de 2015). Censos.
http://mapas.ine.pt/download/index2011.phtml.

[11] Portal da saúde. (21 de Novembro de 2015).
http://www.portaldasaude.pt/portal/servicos/pesquisa/resultados?.

[12] Marble, D. F. 1990. Geographic information systems: an overview. In D. J. Peuquet, & D. F. Marble (editors), Introductory Readings in Geographic Information Systems. London: Taylor & Francis. pp. 8-17.

[13] Open Street Maps (3 de Março de 2016)
http://download.geofabrik.de/europe.html

[14] OpenStreetMap: an alternative to Google Maps (18 Abril 2016)
http://www.webilop.com/openstreetmap-an-alternative-to-google-maps/

[15] Direção-Geral do território portugues (18 de Abril de 2016)
http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_ administrativa_oficial_de_portugal__caop_/caop__download_/

[16] Jornal de negócios (8 de Julho de 2016)
http://www.jornaldenegocios.pt/economia/saude/detalhe/esperanca_de_vida_e_maior_para_quem_tem_maior_ escolaridade_e_rendimento.html

[17] I will teach you (5 de Agosto de 2016)
http://www.iwillteachyoutoberich.com/blog/considering-a-career-in-consulting-avoid-these-5-stupid-mistakes/

[18] Deloitte (10 de Agosto de 2016)
http://www2.deloitte.com/pt/pt.html

5 Anexo I - Algoritmo de comparação de *strings*

Comparar duas *strings* quando estas não são exatamente iguais pode ser uma tarefa árdua quando executada de forma manual. Se pensarmos em comparar centenas de textos é então necessário desenvolver algum algoritmo que seja responsável por esta tarefa e que a consiga executar de forma rápida e eficaz. Tendo em conta que no decorrer deste trabalho será necessário efetuar esta tarefa muitas vezes foram então estudados alguns dos métodos existentes. Na secção presente são descritos alguns métodos de comparação de *strings*.

5.1 Distância de Hamming

A distância de Hamming foi sugerida em 1983 por Sankoff and Kruskal e é usada apenas em casos nos quais os dois textos apresentam o mesmo comprimento correspondendo ao número de caracteres em que estes diferem. Se compararmos por exemplo "rua norte" e "rua forte", a distância de Hamming neste caso é 1, pois as duas moradas diferem apenas em um caracter.

5.2 Método de Ratcliff/Obershelp

O método de Ratcliff/Obershelp ou método de procura de padrões coincidentes foi introduzido em 1983 por John W. Ratcliff e por John A. Obershelp. Segundo este método, obtemos um valor entre 0 e 1, no qual a unidade corresponde a textos iguais e o valor nulo a textos completamente diferentes. A fórmula para o cálculo da distância utilizada neste caso é dada por:

$$D_{R,O} = \frac{2 \times K_m}{|S_1| + |S_2|}$$

na qual K_m representa o número de caracteres coincidentes e $|S_1|$e $|S_2|$ os comprimentos de cada um dos textos. Para definir o número de caracteres coincidentes começamos por procurar a maior sequência coincidente e depois a segunda até não existirem mais coincidências. A seguir está apresentado o exemplo ilustrativo deste tipo de cálculo.

No exemplo da tabela 5.1, a maior igualdade é 'ruad', depois temos 'tome' e por fim 's' pelo que o calculo fica: $D_{R,O} = \frac{2 \times (4+4+1)}{11+12} = 0.78$

O Python disponibiliza uma função com o nome *difflib* que utiliza este método para calcular então a percentagem de coincidência entre os textos.

r	u	a	d	e	s	a	o	t	o	m	e
r	u	a	d	o	s	t	o	m	e	s	

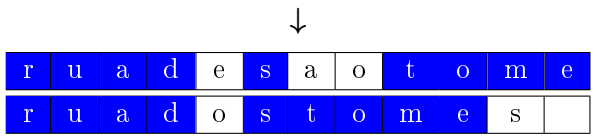


Tabela 5.1: Tabela com os dois nomes a comparar.

5.3 Método de Jaro-Winkler

Introduzido em 1989 por Matthew A. Jaro's, este algoritmo foi criado inicialmente para ser utilizado em dados de saúde. Foi posteriormente alterado por William E. Winkler, que acreditava que a percentagem de similaridade entre dois textos que contêm um grande conjunto de caracteres em comum no início deve ser maior do que quando contêm igualdades nos seguintes. O valor obtido neste caso varia entre 0 e 1 sendo que 1 corresponde a uma coincidência total. A distância era inicialmente dada por:

$$D_J = \frac{1}{3} \times \left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m} \right)$$

em que m é o número de caracteres coincidentes, t o número de caracteres que aparecem trocados e $|S_1|$ e $|S_2|$, os comprimentos de cada um dos textos.

Partindo do mesmo exemplo que anteriormente fica então, $D_J = \frac{1}{3} \times \left(\frac{9}{12} + \frac{9}{11} + \frac{9-0}{9} \right) = 0.86$

Posteriormente foi então alterada para a seguinte forma:

$$D_{JW} = D_J + l \times p \times (1 - D_J)$$

onde l é o número de caracteres coincidentes no inicio da palavra podendo ter um máximo de 4 caracteres e $p \epsilon [0, 0.25]$é um fator de estabilidade.

Usa-se por conveniência o valor de 0,1 para p, sendo neste caso as quatro primeiras letras iguais, l toma o valor 4 ficando a distância então, $D_{JW} = 0.86 + 4 \times 0.1 \times (1 - 0.86) = 0.92$

5.4 Distância de Levenshtein

A distância de Levenshtein ou *edit distance* apareceu pela primeira vez em 1965, definida por Vladimir Levenshtein. Pode ser explicada como o número mínimo de operações para tornar duas palavras iguais, entendendo por operações os processos de inserção, eliminação ou substituição de caracteres. Cada uma destas operações pode ter um custo igual ou podemos definir diferentes custos para cada uma destas.

Para uma melhor explicação deste método considera-se S_1 e S_2 que correspondem a dois textos, sendo o alfabeto representado por $\sum$ e como tal, $S_1, S_2 \, \epsilon \sum$. Os respetivos comprimentos são dados por $|S_1|$ e $|S_2|$. Sendo $D(S_1, S_2)$ a distância de Levenshtein, a forma mais eficiente de obter este valor com o auxílio de programação prende-se com o preenchimento de uma matriz de dimensão $m \times n$, com $m = |S_1|$e $n = |S_2|$. Os elementos da matriz podem ser escritos como:

$$D(S_1, S_2) = \begin{cases} D_{i,0} = i \\ D_{0,j} = j \\ min \begin{cases} D_{i-1,j-1} + 0 & iguais \\ D_{i-1,j-1} + 1 & substituir \\ D_{i-1,j} + 1 & eliminar \\ D_{i,j-1} + 1 & inserir \end{cases} \end{cases}$$

Para mostrar como a tabela é preenchida temos então a Figura 5.1.

		r	u	a	d	a	f	l	o	r
	0	1	2	3	4	5	6	7	8	9
r	1	0	1	2	3	4	5	6	7	7
d	2	1	2	3	2	3	4	5	6	7
a	3	2	2	2	3	2	3	4	5	6
f	4	3	3	3	3	3	2	3	4	5
l	5	4	4	4	4	4	3	2	3	4
o	6	5	5	5	5	5	4	3	2	3
r	7	5	6	6	6	6	5	4	3	2

Figura 5.1: Tabela de cálculo da distância de Levenshtein entre a palavra "ruadesaotome" e "ruadostomes". A distÂncia de Levenshtein é dada pelo número do canto inferior direito, ou seja neste exemplo esta distância é igual a 2.

Sendo a distância de Levenshtein o valor representado no canto inferior direito da tabela, ou seja neste caso corresponde ao valor 4.

No caso da linguagem de programação ser o python, não é necessária a criação de um código tendo em conta que já foram desenvolvidas bibliotecas com esta funcionalidade e se encontram disponíveis online. O nome da biblioteca é Levenshtein e permite calcular o valor de Levenshtein.ratio(S_1,S_2), sendo este valor uma percentagem entre zero e um que utiliza o valor da distância de Levenshtein e o comprimento do maior texto para assim retornar o valor 1 no caso de textos completamente coincidentes e próximo de zero no caso destes serem muito distintos.

Tendo em conta que a distância de Hamming só se adequa a textos com o mesmo comprimento foi automaticamente excluída. Quanto ao método de Jaro-Winkler tem a particularidade de valorizar a igualdade inicial da palavra pelo que se existir uma rua/travessa irá ser muito desvalorizada e se começarem pelo mesmo esta irá ser sobrevalorizada. Este último método não está desenvolvido em nenhuma biblioteca do python o que iria dificultar o trabalho. Assim sendo este é também excluído.

Por fim, considerando o método de Ratcliff/Obershelp e a distância de Levenshtein, ambos com fácil implementação em Python que como vimos anteriormente será a linguagem de programação a utilizar sempre que necessário no decorrer do trabalho. Foram efetuados

testes de performance e precisão para comparação de ambos os métodos e concluiu-se que o primeiro apresenta um consumo de tempo acrescido face ao segundo e tendo em conta que iremos trabalhar com listas de moradas bastante extensas, é mais conveniente utilizar a distância de Levenshtein, sendo este um método eficiente e uma forma rápido para obter as informações necessárias.